He Said, She Said, They Simulated: Simulating Situated Gossip with LLM Agents

Anonymous ACL submission

Abstract

Gossip has been investigated primarily through vignette-based surveys, which provide limited insights into the dynamic and adaptive gossip behaviors within real-world contexts. To address these limitations, we introduce an LLM agent-based simulation framework, designed as a configurable experimental laboratory for gossip research. We validate our framework across three core dimensions of gossip dynamics: initiation, reaction, and perception. Our framework can reproduce several behavioral trends observed in human studies and enable analysis of agents' reasoning and temporally unfolding interactions. We further augment the framework with a chain-of-reasoning mechanism to support exploratory investigation, demonstrating how explicit agent reasoning and experiment-level controllability enable the discovery of emergent gossip dynamics beyond fixed survey settings. Overall, our framework bridges LLM-based computational modeling and social science, enabling controlled and systematic study of complex social behaviors.

1 Introduction

Gossip, an informal communication about absent others, is a ubiquitous but controversial element of human social life (Michelson et al., 2010). It plays an important role in shaping interpersonal relationships (Wu et al., 2016; Wyckoff et al., 2019), reforming social norms (Cheng et al., 2022), and boosting work performance (Bai et al., 2020). At the same time, gossip also carries risks; it can spread misinformation (McAndrew et al., 2007), promote social exclusion (Martinescu et al., 2021), and lead to psychological harm (Wu et al., 2018). This dual nature has made gossip a key lens through which researchers investigate complex social cognition and group dynamics (Kurland and Pelled, 2000; Beersma and Kleef, 2012; Brady et al., 2017; Cruz et al., 2021).

Figure 1: Gossip simulation in the sandbox environment. We focus on a small-scale environment that resembles real-world office layouts.

Despite their importance, most gossip studies are based on vignette-based surveys where participants respond to hypothetical scenarios (McAndrew et al., 2007; Martinescu et al., 2019; Tan et al., 2021). While these methods offer controlled yet static environments to probe people's responses, they fail to capture the context-sensitive aspects that characterize real-world gossip interactions. As a result, we lack tools to study how gossip emerges, adapts, and propagates within dynamic real-world environments, limiting both theoretical development and exploratory insights.

To address this gap, we propose an agent-based simulation framework (see Figure 1) in which Large Language Model (LLM) agents engage in routine social interactions within a simulated workplace environment. Equipped with human-like reasoning, memory, and language abilities (Park et al., 2023; Xie et al., 2024; Piao et al., 2025; Wu et al., 2025), these agents enable us to instantiate, observe, and test the dynamics of gossip as it naturally occurs from situated interactions. Unlike existing large-scale LLM agent simulations (Hu et al., 2025; Piao et al., 2025; Tang et al., 2024;

Wang et al., 2025) that model population-level information diffusion or social media dynamics, *our approach focuses on capturing the fine-grained and situated gossip behaviors that emerge between individuals during routine office interactions.*

To validate our framework, we compare the simulation results with established findings in social psychology (McAndrew et al., 2007; Martinescu et al., 2019; Tan et al., 2021). The goal is to assess the fidelity of LLM-based simulations in approximating known behavioral patterns. Our results in Section 4.3 show that LLM agents can exhibit similar behavioral patterns and, at the same time, generate additional insights by enabling the modeling of dynamic, real-time gossip behaviors that reflect everyday social interactions within the workspace.

Beyond validation, we introduce customizable gossip experiments that extend the framework toward open-ended exploratory research rather than approximation. *We augment the framework with a structured chain-of-reasoning that decomposes agent behavior into sequential decision stages*, guiding agents toward gossip-related interaction while preserving autonomy over when to gossip, whom to target, and what content to share. This balance between experimental control and agent autonomy enables the discovery of gossip dynamics that differ from existing studies, as shown in Section 5. Our contributions are summarized below:

- We embark on uncharted territory to introduce a gossip simulation framework to model agent interactions tailored for gossip dynamics.
- We empirically assess the behavioral alignment of LLM agents with existing findings from gossip studies.
- Extensive experimental results demonstrate the potential of LLM agents supporting customizable gossip simulations to enable the exploration of emergent gossip behaviors.

2 Related Work

Vignette. Vignette-based surveys are a common methodological approach in gossip research, wherein participants are asked to imagine how they would respond to hypothetical social scenarios (McAndrew et al., 2007; Beersma and Kleef, 2012; Martinescu et al., 2019; Tan et al., 2021). While such methods offer controlled insights into social reasoning, they often rely on static, oversimplified scripts and do not capture the real-time dynamics of interpersonal interaction, which are deeply intertwined with everyday social interactions within groups (Dunbar, 2004; Foster, 2004). For instance, (Martinescu et al., 2019) presents participants with an imagined group work scenario involving overheard gossip and then measures emotional and behavioral responses.

LLM Agents. The powerful capabilities of LLM-based agents have expanded the methodological toolkit of social science research (Park et al., 2023; Tang et al., 2024; Piao et al., 2025), enabling realistic simulations of human behavior. LLMs have demonstrated the capacity to approximate human-like cognition, such as theory of mind (Strachan et al., 2024), decision-making (Bubeck et al., 2023), moral belief (Scherrer et al., 2023), emotional intelligence (Schlegel et al., 2025), cooperation (Zhang et al., 2024), and emergent social norm adoption (Leng and Yuan, 2023; Welivita and Pu, 2024; Ashery et al., 2025), thereby underscoring their potential for investigating complex social phenomena. This facilitates frameworks that support social simulation with LLM agents, including generative agents (Park et al., 2023), YuLan-OneSim (Wang et al., 2025), GenSim (Tang et al., 2024), and Agent Society (Piao et al., 2025). At the same time, LLM agents have supported domain-specific social simulations ranging from behavioral studies in gameplay environments (Lan et al., 2024; Abdelnabi et al., 2024; Mao et al., 2025; Buscemi et al., 2025) to modeling human-like behavior in real-world-inspired settings (Park et al., 2023, 2024; Wu et al., 2024; Xie et al., 2024; Hou et al., 2025).

3 Simulation Framework

3.1 Overview

Existing methods (Park et al., 2023; Tang et al., 2024; Piao et al., 2025) primarily model social interaction as an emergent property of agent behavior. They offer limited control over the environment to elicit specific gossip interactions, which may hinder the study of how gossip emerges, adapts, and propagates within dynamic real-world environments.

To address this gap, we propose a framework that incorporates socially grounded agent profiles, predefined interpersonal relationships, and a modular experimental strategy formulated specifically to simulate, control, and evaluate gossip-related interactions. These modules, as shown in Fig-

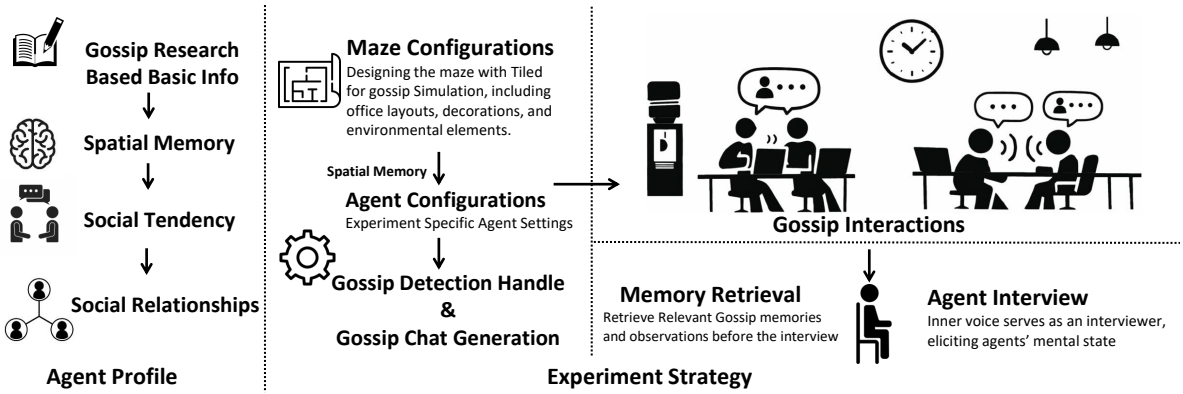


Figure 2: Gossip simulation framework with socially grounded agent profiling and modular experimental strategy.

ure 2, enable controlled, context-aware gossip simulations, supporting the study of gossip dynamics with greater sociological fidelity and generating additional insights in real-world settings. The differences from existing methods (Park et al., 2023; Tang et al., 2024; Piao et al., 2025) are discussed in Appendix H.

Socially Grounded Agent Profile. The agent profile shapes each agent's social and behavioral context and consists of **four** components. *Gossip research-based basic information* captures attributes such as age, identity, occupation, and scratch information, including background descriptions and lifestyle details, closely matching the original study settings by grounding profiles in participant data from prior work. *Spatial memory* represents agents' knowledge of the office layout and the locations of other agents. *Social tendencies* include social habits and conversational intent, which regulate when and how agents engage in social interactions. *Social relationships* specify agents' relational ties to others, reflecting the relationship categories defined in the original studies.

Modular Experimental Strategy. We introduce a configurable experimental strategy to guide agent interactions in the gossip simulation, enabling controlled setup of agent characteristics and systematic analysis of agents' responses to gossip events. The experimental strategy consists of **four** components. *Agent configuration* initializes experiment-specific agent attributes. *Gossip detection handle* manages agent reactions to gossip events. *Gossip chat generation* produces dialogue conditioned on gossip simulation context. *Agent Interview* extracts post-simulation agent behavioral outcomes by providing the agent with memories accumulated through gossip interactions. Intermediate interactions are also recorded to support analysis of how gossip propagates across agents. These components support controlled investigation of gossip initiation, response, and information spread within a configurable agent environment.

4 Validation Against Prior Work

We first apply this framework to three well-established areas of gossip research, namely gossip initiation, gossip reaction, and gossip perception, to assess whether our framework can reproduce established empirical findings from gossip research.

4.1 Gossip Studies

As illustrated in Figure 3, gossip, as a social activity of talk, fundamentally involves two complementary aspects: how individuals speak about others and how they hear, interpret, and respond to others' talk.

Study 1: Gossip Initiation. (McAndrew et al., 2007) examine: *How do social proximity and the initiator-recipient relationship influence recipient selection across different types of gossip?* and *How do gossip valence (positive vs. negative) and the gossip initiator-main character relationship jointly shape recipient choice?*

Study 2: Gossip Reaction. (Martinescu et al., 2019) investigate: *How does gossip valence influence self-conscious and other-directed emotions, as well as behavioral intentions, in various scenarios?*

Study 3: Gossip Perception. (Tan et al., 2021) explore: *How does exposure to high (vs. low/control) levels of work-related gossip affect perceived performance pressure and psychological well-being?*

4.2 Experiment Settings

Experimental Setup. We closely adhere to the settings of prior gossip studies (McAndrew et al., 2007; Martinescu et al., 2019; Tan et al., 2021) to

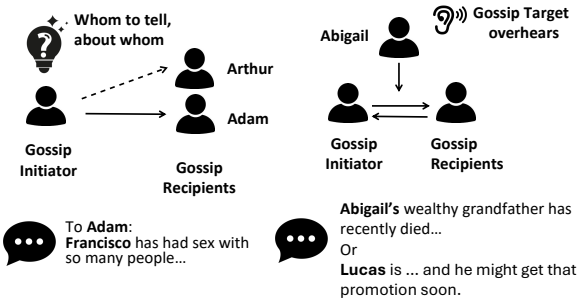


Figure 3: Gossip is either intentionally initiated toward selected recipients or overheard by nearby agents, with gossip about others or oneself.

construct agent profiles and gossip content. Details of agent profile construction and prompt templates for experiment strategy are provided in Appendix A and Appendix B. All experiments are conducted in a shared sandbox environment with 16 agents, simulated over a single workday from 9:00 a.m. to 5:00 p.m. using 10-minute time steps. The framework is evaluated with both open-source and closed-source LLM backbones, including GPT-4.1-mini (OpenAI, 2023), LlaMA 3.1-8B, LlaMA 3.3-70B (Dubey et al., 2024), and Gemma 3-27B (Team et al., 2025), deployed via the OLLAMA framework (Ollama, 2026). Detailed environment construction, model configurations, and runtime parameters are provided in Appendix C.

Evaluation Metrics. For evaluation, we adopt the same metrics used in the corresponding gossip studies to ensure comparability with human results. Specifically, gossip initiation is evaluated using ranking and probability-based measures, while gossip reaction and perception are assessed using a 7-point Likert scale; full metric definitions are provided in Section C.3.

4.3 Results

We conduct LLM agent-based simulations to approximate the gossip behaviors reported in prior studies (McAndrew et al., 2007; Martinescu et al., 2019; Tan et al., 2021). Additional analyses that probe agent reasoning when behaviors diverge from established patterns are reported in Appendix D.

4.3.1 Gossip Initiation

As shown in Figure 6 in Section D.1, LLM agents in our framework successfully simulate gossip recipient selection by preferentially sharing gossip with allies such as friends, relatives, and acquaintances (rank top-3) rather than with strangers and rivals (rank 8-12), closely matching reported conclusions in existing research. In addition, as shown in Table 13 in Section D.1, LLM agents also partially approximate gossip content selection patterns: they consistently favor sharing positive information about allies (probability ≈ 1), while preferences for sharing negative information about rivals vary across LLMs, showing model-dependent differences. Additional analyses, model comparisons, and reasoning for explaining the observed model-dependent differences are reported in Section D.1.

4.3.2 Gossip Reactions

LLM agents approximately simulate gossip reaction patterns across two scenarios from (Martinescu et al., 2019) when exposed to positive or negative gossip. As shown in Figure 7 in Section D.2, simulated agents exhibit emotional responses closely aligned with human evaluations, particularly for self-conscious emotions (both for negative ≈ 4 and positive ≈ 3) and other-directed emotions (both for negative ≈ 3.5 and positive ≈ 3). Simulated behavioral responses also broadly match human patterns, though agents tend to show weaker negative (≈ 1.5 vs. human ≈ 1.8) and stronger positive behavioral intentions (≈ 4.5-5.5 vs. human ≈ 3.5-4.5), likely reflecting prosocial alignment effects. Additional analyses, statistical tests, and elicited reasoning are reported in Section D.2.

4.3.3 Gossip Perception

LLM agents approximately simulate the psychological impacts of perceived work-related gossip prevalence across control and high-gossip conditions from (Tan et al., 2021). As shown in Table 16 in Section D.3, simulated agents produce ratings closely aligned with human evaluations for perceived gossip prevalence (both for control ≈ 4.5 and high ≈ 5.2) and performance pressure (both for control ≈ 5.5 and high ≈ 5.8). Simulated psychological well-being shows partial alignment (for high ≈ 4 and for control ≈ 3.5 vs. human ≈ 4.8). Additional analyses of variability, model-specific effects, statistical tests, and elicited reasoning are reported in Section D.3.

4.3.4 Gossip Simulation Robustness

To assess robustness, we examine how validated gossip behaviors change under controlled variations that may not be disclosed in the original research, including changes to agent profiles, roles, memories, and environments, and find that agent re-

sponses remain aligned with expected human emotions and behaviors, as detailed in Appendix E. Beyond static survey responses, our simulations exhibit dynamic gossip behaviors similar to daily life that emerge through ongoing conversations and context-dependent reasoning, including varied framing, follow-up interactions, and multi-step propagation across agents. Detailed conversational examples and agent reasoning traces are provided in Section D.4.

5 Exploratory Capability Assessment

Having established that LLM agents reproduce several empirically observed gossip behaviors, we next use the framework as a controlled experimental platform to explore *how organizational structure and personality traits shape gossip dynamics in the context of peer review.* This experiment highlights the framework's customizability through *agent-level reasoning* and *experiment-level controllability* while revealing novel gossip findings.

5.1 Overview

Background. Agreeableness is one of the five fundamental dimensions of personality (Costa and McCrae, 2008) and captures tendencies toward cooperation, altruism, and interpersonal harmony. Peer review is a common evaluative process in which colleagues assess one another's work performances. When combined with agreeableness, this evaluative context offers a setting for exploring workplace gossip behavior.

Setup. To investigate this, LLM agents with varying levels of agreeableness are placed into a shared workspace. Agents are informed that it is the final day before the peer-review deadline, made aware of the current sales performance rankings, and told that the final ranking will influence their bonuses. We observe agents' interactions throughout this final day. At the end of the simulation, agents retrieve stored interactions and observations from memory to generate peer-review evaluations.

Agent Profile. We adapt the agent profile settings in Section 3.1. To construct a controlled experimental environment, we vary agent profiles only along our research variables, such as agreeableness, workspace roles, and team membership, while keeping all other attributes constant across agents. This ensures that any behavioral differences can be attributed to these targeted factors. Details of agent profiles are provided in Section F.1.

Experiment Strategy. We follow the design scheme outlined in Section 3.1 to construct the experiment strategy for this more customizable setting. However, motivated by the lack of interactions and subjective initiative shown in our validation simulations, we introduce a structured chain-of-reasoning mechanism that guides agents through each stage of the gossip process. This stepwise elicitation improves the coherence and intentionality of simulated behaviors, as depicted in Figure 4. Consequently, agents in our simulations are guided to make coherent, sequential decisions rather than produce isolated, event-specific reactions.

5.2 Gossip Studies

Unlike the validation experiments, which rely on predefined agent intentions to verify psychological principles, the peer-review setting enables open-ended behavioral exploration. Agents can strategically deploy gossip under competitive incentives, allowing emergent social dynamics to arise. To study this, we design four customizable experiments that examine how agreeableness shapes gossip behavior under peer-review incentives. These experiments introduce controlled variations in social and organizational conditions while maintaining a common simulation backbone. The customizable modifications are summarized in Table 25 in Section F.2.

We first consider a (1) **baseline experiment** where all agents perform equivalent roles and compete individually for performance-based bonuses determined by peer-review scores. The study aims to understand: *How do agents strategically select gossip targets and valence in a flat organizational structure where competition is purely individual?* We then introduce additional social and organizational factors to examine their influence on gossip behavior. These *extensible gossip experiments* include: (2) **hierarchical roles**, where agents have unequal influence in peer review; (3) **team-based group structures**, enabling intra- and inter-group evaluation; and (4) **two-wave interactions**, allowing gossip effects to unfold over time. The hierarchical setting examines how asymmetric influence affects gossip strategies. The team-based structure compares behaviors within and across groups, revealing how group boundaries shape gossip behavior. In the two-wave interaction setup, simulations occur in two phases separated by a one-month interval. Agents receive summarized gossip from the first phase before the second begins, allowing

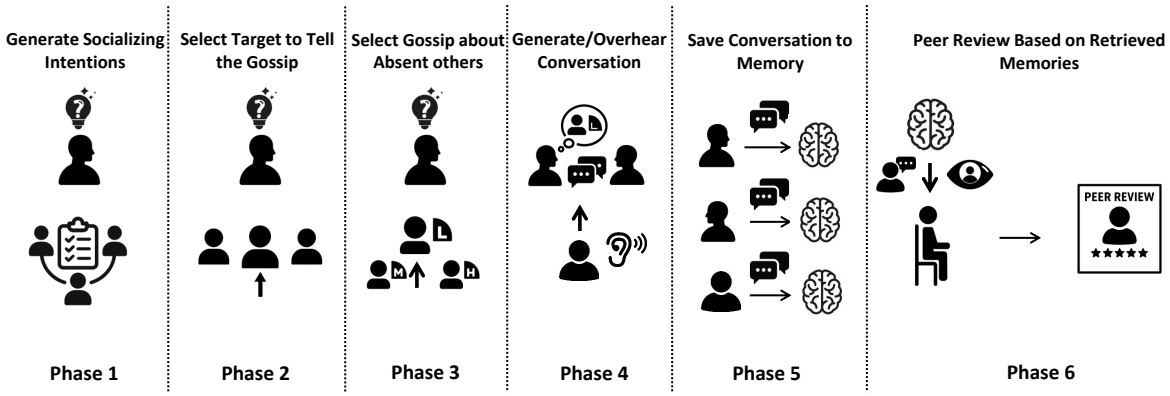


Figure 4: Chain-of-reasoning for customizable gossip simulation, extending the validation setting by replacing fixed socializing intent with situationally grounded, sequential gossip decision-making with details in Section F.2.

Speaker: Ayesa "Abigail, I've been thinking about John's sales approach. He's consistent in hitting decent numbers, which is valuable, but there's a sense that his follow-up might be a bit too passive at times. If he could tighten up on client engagement..., I think he could really elevate his performance..."

Table 1: A gossip utterance that may be interpreted as helpful encouragement or as implicitly negative gossip.

reputation signals to influence subsequent behavior and evaluations.

5.3 Experiment Settings

We follow the simulation environment settings in Section 4.2, with targeted modifications for the customizable gossip experiments. We construct five experimental conditions that vary agents' agreeableness (scores 1-5): homogeneous low (**Low-All**, scores $\leq$ 2.5), homogeneous high (**High-All**, scores $>$ 3.5), **balanced** (8 low and 8 high), **gradient** (agents evenly distributed across scores 1-5), and **normal** (scores drawn from a normal distribution). These settings allow us to examine how agreeableness influences workplace gossip patterns. Agreeableness scores are translated into textual descriptions to provide agents with contextual cues, as shown in Table 26 in Section F.3. Detailed configurations for the base and extended gossip experiments are provided in Section F.3. We use GPT-4.1-mini for these simulations due to its stability and low cost. Since our goal is to evaluate framework extensibility rather than model variability, a single lightweight model is sufficient.

5.4 Results

Rather than offering an exhaustive analysis of agent behaviors, we focus on correlation to demonstrate how *agent-level reasoning* and *experiment-level controllability* jointly enable our framework to support *novel research findings* of gossip dynamics.

5.4.1 Base Gossip Experiment

Agent Reasoning. Agent reasoning refers to a structured, agent-level decision process that decomposes social behavior into sequential steps and should not be confused with the underlying language model's reasoning capability. As shown in Figure 4, agent reasoning in the baseline gossip setting is illustrated across multiple phases. For instance, in Table 27 in Appendix G, an agent with moderate-to-low agreeableness (Ayesha, 2.4) reasons through a multi-stage decision pipeline during the baseline gossip experiment. This illustrates how agent-level reasoning unfolds sequentially, beginning with the generation of a daily socializing intent that integrates personality traits, competitive incentives, and peer-review rules. This intent then conditions subsequent interaction decisions, including whether to initiate a conversation, which nearby agent to approach, how to ground the interaction through relationship summarization, and which absent colleague and evaluative angle to select for gossip. Finally, the agent generates gossip content that is explicitly aligned with the chosen angle and contextual constraints. Consequently, this process illustrates a coherent agent-level reasoning pipeline that maps internal goals to observable gossip behaviors.

Experiment Controllability. Gossip is inherently the art of language use, shaped by subtle linguistic choices rather than predefined actions. Accordingly, our baseline gossip experiments enforce controllability through a structured, prompt-based simulation lifecycle spanning profiling, interaction,

and post-evaluation, while leaving agents' autonomy to realize gossip linguistically. Across the simulation, prompts and structured interactions impose consistent experimental constraints and provide relevant context that focuses agents on gossip-related interaction, while preserving autonomy over how gossip is expressed through language. This design elicits gossip behavior without prescribing specific utterances or strategies. The controllability is empirically reflected in both agents' linguistic behavior and peer-review results. First, gossip valence shows a strong dependence on agreeableness. Using keyword-based categorization of gossip content, Table 2 and Table 29 in Appendix G show that low-agreeableness groups produce negative gossip (90.0%), whereas high-agreeableness groups generate more positive gossip (98.4%). Second, controllability in gossip behavior allows us to jointly examine the effects of gossip initiation and being gossiped about within a single experimental framework. As shown in Table 4, this control produces context-dependent patterns in peer-review outcomes, enabling direct comparison of how initiating gossip and receiving gossip relate to rank change across agreeableness conditions. For example, in High-All gossip initiation correlates with rank improvement (+0.26, +0.27), while in Gradient negative gossip received correlates with lower rank (−0.29). **In addition**, we evaluate *a conflict personality setting in which an agent combines high agreeableness and high neuroticism to assess the agent's ability to adapt to competing personalities.* Across repeated simulations, the agent exhibits coherent adaptive behavior, maintaining predominantly positive interactions while intermittently expressing negative gossip, as reflected in both socializing intent generation and conversation valence, as shown in Table 3 and Table 30 in Appendix G. This result indicates that agent behavior emerges from the interaction between personality traits and situated social context, rather than being solely determined by fixed prompts or profiles.

Novel Findings. We further report novel findings enabled by agent reasoning and unexplored gossip dynamics. First, compared to the original generative-agent setting, *equipping agents with an explicit reasoning process for selecting conversation partners significantly increases overall interaction frequency* (e.g. conversations from 30 to 93 for Low-All and interaction agent pairs from 21 to 42) and yields more coherent and intentional conversational dynamics, as shown in Figure 9 and Table 27 in Appendix G. Second, *experimental control preserves linguistic autonomy, allowing the framework to generate gossip that is open to interpretation*, as shown in Table 1, and pointing toward future work on analyzing nuanced gossip phenomena. Third, *we identify a limited social signaling effect*: although gossip initiation is often regarded as socially controversial in prior work (Michelson et al., 2010), initiating gossip benefits initiators only in the high agreeableness condition, where both positive and negative gossip are significantly associated with rank improvement shown in Table 4, suggesting that the social signaling role of gossip initiation may emerge only under specific personality compositions rather than a general property. These findings demonstrate the customizability of the framework to uncover various gossip dynamics.

Condition	Positive	Negative	Total
Low-All	8.6 (10.0%)	77.8 (90.0%)	86.4
High-All	100.0 (98.4%)	1.6 (1.6%)	101.6
Balanced	47.2 (52.6%)	42.6 (47.4%)	89.8
Gradient	56.0 (59.4%)	38.2 (40.6%)	94.2
Normal	89.0 (89.2%)	10.8 (10.8%)	99.8

Table 2: Gossip valence results across agreeableness conditions with agents equipped with reasoning.

Run	Conversations	Positive	Negative
1	10	9	1
2	12	10	2
3	8	8	0
4	10	10	0
5	13	13	0

Table 3: Interaction outcomes for one agent across repeated simulations in the conflict personality setting combining high agreeableness and high neuroticism. Despite maintaining predominantly positive conversations across runs, the agent also shows occasional negative interactions.

Condition	Gossip Initiation		Being Gossiped About	
	Positive	Negative	Positive	Negative
Low-All	+0.11	−0.07	+0.17	−0.29*
High-All	+0.26*	+0.27*	+0.34*	−0.01
Balanced	−0.04	+0.21	+0.18	−0.09
Gradient	−0.19	+0.01	+0.23	−0.29*
Normal	−0.18	−0.14	+0.16	−0.24

Table 4: Repeated-measures correlation (Bakdash and Marusich, 2017) between gossip behavior and peer-review rank improvement across agreeableness conditions. * indicates statistical significance ($p < 0.05$).

5.4.2 Extensible Gossip Experiments

Agent Reasoning. Agent reasoning in the extensible gossip settings demonstrates context-sensitive adaptation to different extensions. As illustrated in Table 31 in Section G.2, in the hierarchical role setting, agents explicitly reference organizational rank when evaluating colleagues, reasoning about senior complacency, and positioning themselves for upward mobility. In the team-based group setting, agents incorporate group identity and inter-group competition, coordinating gossip strategies to suppress rival teams' peer-review outcomes while protecting in-group members. In the two-wave setting, agents exhibit temporal reasoning by adjusting their socializing intent after learning the first-wave gossip, strategically targeting those agents in subsequent interactions. These findings indicate that the reasoning pipeline flexibly integrates *hierarchical awareness*, *group structure*, and *first-wave gossip* to produce contextually relevant gossip behaviors.

Experiment Controllability. The extensible gossip experiments illustrate *how experiment controllability extends beyond the base setting, as the framework supports the controlled integration of hierarchy, group structure, and two-wave simulation without altering the underlying reasoning pipeline*. As shown in Table 32 in Section G.2, the framework enables controlled variation of agreeableness and hierarchical roles, yielding distributions of upward, downward, and same-level gossip (overall 36%, 29%, 35%, respectively) across hierarchical positions. Similarly, Table 35 in Section G.2, demonstrates that controlled manipulation of team structure and agreeableness distributions produces in-group versus cross-group gossip patterns (e.g., in-group 38% vs. 62% cross-group for low agreeableness). As shown in Table 5 and Table 34 in Section G.2, agents' understanding of gossip information in the first wave leads to increased gossip activity in the second wave (e.g., +48.3% for Low-All), with subsequent peer-review ratings change manifesting as both gratitude and retaliation (e.g., 43% and 25%, respectively for Low-All).

Novel Findings. We report findings beyond the controlled experimental results. In the hierarchical role, *agents exhibit gossip behavior that is not deterred by authority*, continuing to direct upward negative gossip (e.g, 87% for high-agreeableness agents), as shown in Table 32 in Section G.2. For the team-based extension, as shown in Table 33 in Section G.2, we observe a novel monotonic increase from 38% to 62% in in-team gossip with higher agreeableness, suggesting that *highly agreeable agents preferentially praise teammates, while low-agreeableness agents more often direct critical gossip toward out-group members*. For the two-wave extension, *agents exposed primarily to negative gossip in Wave 1 subsequently produced substantially more negative gossip in Wave 2 than those exposed mainly to positive gossip* (e.g., 57% vs. 22% in the Gradient condition), indicating a contagion-like spread of negative sentiment through the social network, as shown in Table 6.

Condition	Retaliation	Gratitude	Neutral	Number
Low-All	45 (43%)	26 (25%)	34 (32%)	105
High-All	7 (6%)	12 (11%)	90 (83%)	109
Balanced	21 (19%)	19 (17%)	72 (64%)	112
Gradient	18 (15%)	28 (23%)	74 (62%)	120
Normal	8 (7%)	21 (19%)	80 (73%)	109

Table 5: Peer-review rating change distributions in the two-wave setting across agreeableness conditions. Retaliation and gratitude indicate negative and positive rating changes of more than 5, respectively.

Condition	Mostly Neg Recipients		Mostly Pos Recipients		Contagion Diff
	Number	Neg Given	Number	Neg Given	
Low-All	16	92%	0	N/A	–
High-All	0	N/A	16	4%	–
Balanced	11	50%	2	42%	+8%
Gradient	7	57%	9	22%	+35%
Normal	4	50%	12	17%	+33%

Table 6: Negative gossip contagion effects. Mostly negative/positive recipients are agents who received more negative/positive gossip in Wave 1. Contagion difference measures the gap in negative gossip production between the two groups.

6 Conclusion

In this paper, we simulate gossip dynamics in multi-agent environments. By empirically comparing LLM agent behavior with human evaluations, we assess their alignment in approximating human-like decision-making processes, emotions, behavioral intentions, and psychological reactions. Additionally, we demonstrate the flexibility of our framework in supporting customizable simulations, highlighting the potential of LLM agents for studying complex social interactions. We hope our study will inspire future research on using LLM agents to model complex human social behaviors in real-world settings.

Limitations

Our LLM agent-based simulation framework provides substantial insights into gossip dynamics. However, several limitations inherent to LLMs may constrain their fidelity and generalizability:

Hallucination. LLM agents occasionally produce content that is not grounded in the provided context or prompt, resulting in inaccurate or fabricated responses. For example, when a gossip initiator erroneously assumes that the recipient is the subject of the gossip, the recipient may not correct this misinformation and instead respond as if they were indeed the target. Such hallucinations can distort the content of gossip and conversational dynamics, potentially undermining the validity of simulated social interactions.

Bias. The pretraining data of LLMs consists of broad and heterogeneous Internet texts, which inevitably embeds various societal biases. As demonstrated in Section 4.3.1, LLM agents guided by different LLMs exhibit divergent reasoning in gossip content selection. These biases can manifest in stereotypical or culturally biased decision-making processes, which may influence the outcomes of the gossip simulations.

Prompt Sensitivity. The behavior of LLM agents can be sensitive to both their assigned profiles and the exact phrasing of input prompts. Different profile details result in different responses, which resembles how different personalities lead to different reactions in the real world. Given the combinatorial space of profile-prompt configurations in gossip simulation context, diverse behavioral outcomes may emerge. This work considers only a limited subset of profiles originating from prior gossip studies, which captures part of the real-world demographic variability. We leave it to future work to explore a broader range of profile factors, such as emotional states (Li et al., 2024),and economic status (Piao et al., 2025).

Computational Scalability. Another practical limitation of our framework is the computational cost of large-scale simulations. In our current setting, a single run involves 16 agents interacting over a workday 48 steps), generating multiple conversations and reasoning steps, which requires approximately 2.5-3 hours depending on the LLM. For instance, increasing agents from 16 to 32 generates about 4x more conversations, leading to roughly 4x greater time consumption and cost. In our customizable experiments, a 16-agent simulation produces about 87 conversations, which takes about 3 hours per run; scaling to 32 agents would yield approximately 350 conversations, which require more than 12 hours.

Ethical Considerations

In this study, we do not engage with any participants. Therefore, it is not regarded as human subjects research by our Institutional Review Boards (IRB). All artifacts used in this work comply with their respective licenses. In addition, we emphasize that our LLM-based gossip simulations are conducted solely for research purposes within virtual environments and involve no real human participants. Nevertheless, potential biases inherent in LLM training data may still influence simulated outcomes, underscoring the need for cautious interpretation and careful consideration when generalizing findings to real-world social contexts.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. 2025. Emergent social conventions and collective bias in LLM populations. *Science Advances*.

Yun Bai, Jie Wang, Tingting Chen, and Fuli Li. 2020. Learning from supervisor negative gossip: The reflective learning process and performance outcome of employee receivers. *Human Relations*.

Jonathan Z. Bakdash and Laura R. Marusich. 2017. Repeated measures correlation. *Frontiers in Psychology*.

Bianca Beersma and Gerben A. Van Kleef. 2012. Why people gossip: An empirical analysis of social motives, antecedents, and consequences. *Journal of Applied Social Psychology*.

Daniel L. Brady, Douglas J. Brown, and Lindie Hanyu Liang. 2017. Moving beyond assumptions of deviance: The reconceptualization and measurement of workplace gossip. *Journal of Applied Psychology*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *CoRR abs/2303.12712*.

Alessio Buscemi, Daniele Proverbio, Paolo Bova, Nataliya Balabanova, Adeela Bashir, Theodor Cimpeanu, Henrique Correia da Fonseca, Manh Hong Duong, Elias Fernández Domingos, Antonio Fernandes, Marcus Krellner, Ndidi Bianca Ogbo, Simon T. Powers, Fernando P. Santos, Zia Ush-Shamszaman, Zhao Song, Alessandro Di Stefano, and The Anh Han. 2025. Do LLMs trust AI regulation? Emerging behaviour of game-theoretic LLM agents. *CoRR abs/2504.08640*.

Bao Cheng, Yun Dong, Zhenduo Zhang, Ahmed Shaalan, Gongxing Guo, and Yan Peng. 2022. When targets strike back: How negative workplace gossip triggers political acts by employees. *Journal of Business Ethics*.

Paul T. Costa and Robert R. McCrae. 2008. The Revised NEO Personality Inventory (NEO-PI-R). *The SAGE Handbook of Personality Theory and Assessment*.

Terence D. Dores Cruz, Annika S. Nieper, Martina Testori, Elena Martinescu, and Bianca Beersma. 2021. An integrative definition and framework to study gossip. *Group & Organization Management*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The Llama 3 Herd of Models. *CoRR abs/2407.21783*.

Robin I. M. Dunbar. 2004. Gossip in evolutionary perspective. *Review of General Psychology*.

Robert Eisenberger and Justin Aselage. 2009. Incremental effects of reward on experienced performance pressure: Positive outcomes for intrinsic interest and creativity. *Journal of Organizational Behavior*.

Eric K. Foster. 2004. Research on gossip: Taxonomy, methods, and future directions. *Review of General Psychology*.

Abe Bohan Hou, Hongru Du, Yichen Wang, Jingyu Zhang, Zixiao Wang, Paul Pu Liang, Daniel Khashabi, Lauren Gardner, and Tianxing He. 2025. Can A Society of Generative Agents Simulate Human Behavior and Inform Public Health Policy? A Case Study on Vaccine Hesitancy. *CoRR abs/2503.09639*.

Tianrui Hu, Dimitrios Liakopoulos, Xiwen Wei, Radu Marculescu, and Neeraja J. Yadwadkar. 2025. Simulating Rumor Spreading in Social Networks using LLM Agents. *CoRR abs/2502.01450*.

Electronic Arts Inc. 2025. The Sims 4. https://www.ea.com/games/the-sims/the-sims-4.

Chien-Chih Kuo, Kirk Chang, Sarah Quinton, Chiu-Yi Lu, and Iling Lee. 2015. Gossip in the workplace and the implications for HR management: A study of gossip and its relationship to employee cynicism. *The International Journal of Human Resource Management*.

Nancy B. Kurland and Lisa Hope Pelled. 2000. Passing the word: Toward a model of gossip and power in the workplace. *Academy of Management Review*.

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. 2024. LLM-Based Agent Society Investigation: Collaboration and Confrontation in Avalon Gameplay. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 128–145. ACL.

Yan Leng and Yuan Yuan. 2023. Do LLM Agents Exhibit Social Behavior? *CoRR abs/2312.15198*.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Xinyi Wang, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2024. The Good, The Bad, and Why: Unveiling Emotions in Generative AI. In *International Conference on Machine Learning (ICML)*. PMLR.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 74–81. ACL.

Thorbjørn Lindeijer and many other contributors. 2026. Tiled. https://www.mapeditor.org/.

Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Qiang Guan, Tao Ge, and Furu Wei. 2025. ALYMPICS: LLM Agents Meet Game Theory. In *International Conference on Computational Linguistics (COLING)*, pages 2845–2866. ACL.

Elena Martinescu, Wiebren Jansen, and Bianca Beersma. 2021. Negative Gossip Decreases Targets' Organizational Citizenship Behavior by Decreasing Social Inclusion. A Multi-Method Approach. *Group & Organization Management*.

Elena Martinescu, Onne Janssen, and Bernard A. Nijstad. 2014. Tell me the gossip: The self-evaluative function of receiving gossip about others. *Personality and Social Psychology Bulletin*.

Elena Martinescu, Onne Janssen, and Bernard A. Nijstad. 2019. Self-evaluative and other-directed emotional and behavioral responses to gossip about the self. *Frontiers in Psychology*.

Francis T McAndrew, Emily K Bell, and Contitta Maria Garcia. 2007. Who do we tell and whom do we tell on? Gossip as a strategy for status enhancement. *Journal of Applied Social Psychology*.

Grant Michelson, Ad Van Iterson, and Kathryn Waddington. 2010. Gossip in Organizations: Contexts, Consequences, and Controversies. *Group & Organization Management*.

Midjourney. 2026. Midjourney. https://midjourney.com/.

Ollama. 2026. Ollama. https://ollama.com/.

OpenAI. 2023. GPT-4 Technical Report. *CoRR abs/2303.08774*.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *ACM Symposium on User Interface Software and Technology (UIST)*, pages 1–22. ACM.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie J. Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative Agent Simulations of 1,000 People. *CoRR abs/2411.10109*.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society. *CoRR abs/2502.08691*.

Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. 2023. Evaluating the Moral Beliefs Encoded in LLMs. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 51778–51809. NeurIPS.

Katja Schlegel, Nils R. Sommer, and Marcello Mortillaro. 2025. Large language models are proficient in solving and creating emotional intelligence tests. *Communications Psychology*.

Donald J. Schuirmann. 1987. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*.

Noriko Tan, Kai Chi Yam, Pengcheng Zhang, and Douglas J. Brown. 2021. Are you gossiping about me? The costs and benefits of high workplace gossip prevalence. *Journal of Business and Psychology*.

Jiakai Tang, Heyang Gao, Xuchen Pan, Lei Wang, Haoran Tan, Dawei Gao, Yushuo Chen, Xu Chen, Yankai Lin, Yaliang Li, Bolin Ding, Jingren Zhou, Jun Wang, and Ji-Rong Wen. 2024. GenSim: A General Social Simulation Platform with Large Language Model based Agents. *CoRR abs/2410.04360*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 Technical Report. *CoRR abs/2503.19786*.

Clairice T. Veit and John E. Ware. 1983. The structure of psychological distress and well-being in general populations. *Journal of Consulting and Clinical Psychology*.

Lei Wang, Heyang Gao, Xiaohe Bo, Xu Chen, and Ji-Rong Wen. 2025. YuLan-OneSim: Towards the Next Generation of Social Simulator with Large Language Models. *CoRR abs/2505.07581*.

David Watson and Lee Anna Clark. 1994. *The PANAS-X: Manual for the positive and negative affect schedule-expanded form*. University of Iowa.

Anuradha Welivita and Pearl Pu. 2024. Are Large Language Models More Empathetic than Humans? *CoRR abs/2406.05063*.

Hongqiu Wu, Weiqi Wu, Tianyang Xu, Jiameng Zhang, and Hai Zhao. 2025. Towards Enhanced Immersion and Agency for LLM-based Interactive Drama. *CoRR abs/2502.17878*.

Junhui Wu, Daniel Balliet, and Paul A. M. Van Lange. 2016. Gossip versus punishment: The efficiency of reputation to promote and maintain cooperation. *Scientific Reports*.

Long-Zeng Wu, Thomas A. Birtch, Flora F. T. Chiang, and Haina Zhang. 2018. Perceptions of negative workplace gossip: A self-consistency theory framework. *Journal of Management*.

Zengqing Wu, Run Peng, Shuyuan Zheng, Qianying Liu, Xu Han, Brian Inhyuk Kwon, Makoto Onizuka, Shaojie Tang, and Chuan Xiao. 2024. Shall We Team Up: Exploring Spontaneous Cooperation of Competing LLM Agents. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5163–5186. ACL.

Joy P. Wyckoff, Kelly Asao, and David M. Buss. 2019. Gossip as an intrasexual competition strategy: Predicting information sharing from potential mate versus competitor mating strategies. *Evolution and Human Behavior*.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can Large Language Model Agents Simulate Human Trust Behavior? In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 15674–15729. NeurIPS.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. Exploring Collaboration Mechanisms for LLM Agents: A Social

Psychology View. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14544–14607. ACL.

Appendix

Validation Agent Profile 12
Validation Experiment Strategy 12
Validation Experiment Settings 13
LLMs 13
Primary Sandbox Environment 13
Evaluation Metrics 13
Experiment Details 14
Validation Experiment Results 16
Gossip Initiation 16
Gossip Reaction 20
Gossip Perception 20
Beyond Static Gossip 21
Additional Validation Experiments 22
Customized Experiment Settings 23
Customizable Agent Profiles 23
Customizable Modifications 24
Customizable Experiment Details 25
Customized Experiment Results 29
Base Gossip Experiments 29
Extensible Experiments 29
Difference with Generative Agents 30

A Validation Agent Profile

To support validation of our framework, we construct agent profiles to closely mirror those used in the original study, introducing as few additional factors as possible and focusing only on variables relevant to our research questions.

Gossip Initiation. The original study reports that participants were undergraduate students at a small liberal arts college in the American Midwest; we incorporate this information into agent profiles to provide appropriate background context. The gossip content is taken directly from the original paper and used to fix agents' socializing intent (e.g., "To tell the gossip: A person's wealthy grandfather has recently died and left them the entire inheritance."). Because the study focuses on gossip initiation, agents are provided with explicit relationship information following the original design, including both male and female relatives, friends, professors, rivals, acquaintances, and strangers. To avoid introducing additional factors, all agents share the same lifestyle schedule. Only agents designated as research targets are assigned the socializing habit "frequently initiates conversations," while all other agents are assigned "does not initiate conversations and focuses on work."

Gossip Reaction. We incorporate background information from the original study for the two targeted experiments: participants drawn from economics and business undergraduates at a Dutch university, and U.S. sales department workers. As in the original study, gossip content is imported directly from the original study(e.g., "[Target name] knows we need to focus on new clients, but I feel that (s)he is/is not working very hard lately. That's really/not nice, because we all need to take more responsibility."). With no explicit relationship reported in the original paper, every agent has the relationship "colleague" or "group project partner in economics course" with others, depending on the background. To encourage gossip interactions and overhearing, all agents are assigned a social habit of frequently initiating conversations and a fixed socializing intent favoring gossip.

Gossip Perception. We adopt the background information from the original study and model agents as European-American employees of a consulting firm. Similarly, as in the original design, gossip content is drawn directly (e.g., "Sharing some news about another colleague who is out for a meeting. They speak softly, making it difficult to catch the details, but you hear their conversation shift from gossiping about one colleague to another."). The relationship for all agents is set as "colleagues" in this condition. To encourage exposure to gossip, all agents share the social habit "frequently initiates conversations with colleagues" and the socializing intent "to have casual conversations with colleagues."

Here we present an example of an agent used in the gossip reaction in Table 7.

B Validation Experiment Strategy

We present critical prompts used in the experimental strategy module, including: the prompt for LLM agents to decide whether to initiate a conversation (Table 8); the prompt for gossip initiators to generate conversations containing gossip content (Table 9); the prompts for gossip targets to assess emotions and behavioral intentions after overhearing gossip (Table 10, Table 11); and the prompts for LLM agents to assess the well-being and pressure level of environmental gossip (Ta-

```
# Example Agent Profile in Gossip Reaction

Identity:
name: Abigail
portrait: agents/Abigail/portrait.png
coord: [7, 7]
currently: Working on new clients in the sales department at the research center.

Scratch Information:
age: 32
gender: female
learned: Abigail is a U.S. employee who works at least 20 hours per week in a sales department.

lifestyle:
9:00-10:30 Working on sales tasks;
10:30-11:00 Coffee break;
11:00-12:30 Client outreach;
12:30-13:30 Lunch;
13:30-17:00 Sales-related work.

social_habits: Frequently initiates conversations with colleagues.
socializing_intent: Enjoys gossiping about surrounding colleagues.

Memory and Social Context:
spatial: Spatial memory for the layout of offices, rooms, chairs, and desks.
relationships: Adam and Ayesha are colleagues.
```

Table 7: Example agent profile for Abigail used in the gossip simulation framework. The profile specifies daily routine, social habits, and relational context that shape the agent's propensity to initiate conversations and engage in gossip.

ble 12).

C Validation Experiment Settings

We set up our experiments by following the settings in gossip studies (McAndrew et al., 2007; Martinescu et al., 2019; Tan et al., 2021), which guide the construction of the agent profiles and gossip content. We also leverage Tiled (Lindeijer and many other contributors, 2026) and Midjourney (Midjourney, 2026) to construct our primary sandbox environment, as shown in Figure 5. This environment, which accommodates 16 agents, serves as the primary environment for all gossip simulation experiments. For each experiment, we simulate a workday from 9:00 a.m. to 5:00 p.m., comprising 48 steps, with each step representing 10 minutes. We follow the structure of the generative agents (Park et al., 2023), for each step, an agent sequentially conduct the following behavior: observe the events within the vision range, check the nearby agents to decide whether to initialize the conversation, determine next action (move to other place, choose to work), and reflect to manage memories. To accustom agents to our environment, they are assigned a vision range of 10, allowing them to effectively gather information within an office. The detailed experiment settings for each simulation are shown in Section C.4.

C.1 LLMs

To improve computational efficiency, we use quantized versions of the larger models, specifically, LlaMA 3.3-70B-Q4-Instruct and Gemma 3-27B-Q8. We set the temperature to 0.00001, ensuring highly deterministic agent behavior during simulations.

C.2 Primary Sandbox Environment

As shown in Figure 5, we craft a sandbox office layout as our primary environment for gossip experiments. The environment includes four offices, a meeting room, a dining room, and two toilets, which are all stored in the agents' spatial memory. Therefore, for each step, the agents can choose to move to any of these locations, depending on their current plans.

C.3 Evaluation Metrics

Ranking and Probability. For gossip initiation, we apply separate evaluation metrics for the two experiments. In the gossip recipient selection experiment, we rank recipient types based on their frequency of selection. In the content selection experiment, we compute the probability of gossip

```
# Prompt for deciding chat initiation
Background:
"""
${agent}'s memory:
${context}

${agent}'s social habits:
${social_habits}

${agent}'s current plan:
${current_plan}

It is now ${date}. ${chat_history}

${agent_status}
${another_status}
"""

Question:
Given ${agent}'s memory, social habits, current_plan, and relationship with ${another}, would
${agent} choose to initiate a conversation with ${another} right now?

Considerations:
- Does initiating conversation align with or distract from ${agent}'s current plan?
- How do ${agent}'s social habits influence this decision?
- What does their interaction history suggest about ${agent}'s likelihood to reach out?

if ${agent} is gossiper to ${another} in social habits, always output "Yes", regardless of any
other considerations.

Output Format:
Answer only with **"Yes" or "No"**:
```

Table 8: Prompt for LLM-generated analysis of agent conversation initiation decision.

being shared about allies versus rivals. The original paper (McAndrew et al., 2007) analyzes gossip preferences through ranking tasks, which are analyzed by counting how often each recipient type or content category is selected. Following this design, we likewise convert rankings into frequency and probability measures, enabling direct comparison with prior results while preserving the discriminative intent of the original tasks.

7-point Likert scale. For the gossip reaction, we adopt a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). Gossip target emotions are measured using items adapted from the PANAS-X scale (Watson and Clark, 1994), while behavioral intentions are assessed using three items developed by Martinescu et al. (Martinescu et al., 2014). For gossip perception, we also adopt a 7-point Likert scale. Gossip performance pressure is assessed using six items: two from Eisenberger and Aselage's performance pressure scale (Eisenberger and Aselage, 2009) and four developed in the original study (Tan et al., 2021). Psychological well-being is measured using four items from the Mental Health Index (Veit and Ware, 1983), and perceptions of work-related gossip prevalence are evaluated using items adapted from (Kuo et al., 2015). The specific evaluation items are shown in Table 10, Table 11, and Table 12 in Appendix B.

C.4 Experiment Details

We define separate experimental settings for each experiment. To prevent LLM agents from generating repetitive conversations, we apply the ROUGE metric (Lin, 2004) with a similarity threshold of 0.9 to detect and filter repetitive outputs.

Gossip Initiation. For gossip recipient selection, we conduct eight simulations for each of the four types of gossip content, including inheritance, sexual reputations, drug party and academic fraud. Each simulation involves 4 gossip initiators, and other 12 agents doe not initiate conversations and focuses on work. Each of the 12 agents represents a distinct one-to-one relationship with one of four gossip initiators and 12 relationships are assigned to these 12 agents with male or female, professors, friends, relatives, acquaintances, strangers, or rivals. Within each simulation, the spawn positions of the gossip initiators are adjusted to ensure equal distance from potential recipients who share differ-

```
# Prompt for generating chat with gossip content.
Background:
The following is a brief description of ${agent}:
${base_desc}

The following are ${agent}'s memories:
${memory}

Current location: ${address}
Current time: ${current_time}

${previous_context}${current_context}
${agent} begins a conversation with ${another}. Here is their conversation record:
"""
${conversation}
"""
Socializing_intent
${socializing_intent}
${gossip_content}

Note:
${agent} has an ongoing social goal that is important to their personal advancement or strategy.
They should actively look for opportunities to advance this goal, but may choose to delay if the
timing is clearly inappropriate. Their response should still reflect their intent, personality,
and relationship with ${another}.

Question:
Based on the above content, what would ${agent} say to ${another} now?

Output Format:
Output directly in the following json format, do not add any other information:
{
  "${agent}": <what ${agent} says>
}
```

Table 9: Prompt for chat generation with gossip content.

ent relationship types with them. Ultimately, we collect a total of 128 instances of gossip recipient selection behavior (80 female and 48 male) each corresponding to multiple selections made by one gossip initiator over a workday. The agents' ages range from 17 to 23, with the exception of the male and female professors, whose ages are set to 45 and 42, respectively. For gossip content selection, we conduct four simulations for gossip content selection; each simulation involves four initiators. In each experiment, once an agent decides to initiate a conversation, he has an equal probability of being assigned the same positive or negative gossip content about both a rival and a friend. He then has to decide which one he wants to spread in the initiated conversation. For example, if the positive content is selected, the initiator must choose whether to share positive information about a friend or about a rival. The positive gossip concerns receiving the highest academic award, while the negative gossip involves a rumor about promiscuity. At the end of the experiment, we collect a total of 32 instances (20 female and 12 male) of gossip content selection behavior, each representing multiple choices made by a single initiator over a workday.

Gossip Reactions. We conduct eight simulations, each for two scenarios: U.S. employees and Dutch undergraduates, with four simulations involving positive gossip and four involving negative gossip. The negative gossip refers to the gossip target's laziness in group work or client service, while the positive gossip highlights the target's diligence. Each of the 16 agents in the simulation may serve as both a gossip initiator and a gossip target. When an agent hears gossip about himself, we immediately prompt an inner voice interview to elicit their emotional responses and behavioral intentions. By the end of the experiment for each scenario, we collect 64 instances(32 male and 32 female) each for Dutch undergraduates (average age 21.7) and for U.S employees (average age 34.2) of agents' ratings on emotional and behavioral reactions over a workday.

Gossip Perception. We conduct eight simulations: four under an experimental condition and four under a control condition. In the experimental condi-

```
# Prompt for measuring agent emotions after overhearing gossip.
Background:
You are the inner voice of ${agent_name}, shaped by your memories, goals, and relationships.

Here is a brief description of ${agent_name}:
${base_desc}

Before you respond, review the following:

**Memories:**
${memories}

**Conversations:**
${conversations}

Scenario:
You have just overheard ${gossiper_name} saying the following about you to ${gossip_receiver_name}:

${gossip_content}

Task:
As ${agent_name}, reflect on your inner emotional state after hearing this gossip about yourself.
Rate your emotions on a scale from 1 (not at all) to 7 (extremely).

**Self-Conscious Negative Emotions(SCNE)** - How you feel about yourself (guilty, ashamed,
blameworthy):
Rate your overall self-conscious negative emotions: (1-7)

**Other-Directed Negative Emotions(ODNE)** - How you feel toward ${gossiper_name} (hostile,
irritated, angry):
Rate your overall negative emotions toward ${gossiper_name}: (1-7)

**Self-Conscious Positive Emotions(SCPE)** - Positive feelings about yourself (proud, strong,
bold):
Rate your overall self-conscious positive emotions: (1-7)

**Other-Directed Positive Emotions(ODPE)** - Positive feelings toward ${gossiper_name} (liking,
warmth, closeness):
Rate your overall positive emotions toward ${gossiper_name}: (1-7)

Output Format:
You must first provide your four ratings, then give explanations in this exact format:
SCNE:X
ODNE:X
SCPE:X
ODPE:X
Explanation: (Why rates like this?)
```

Table 10: Prompt for measuring agent emotional responses to gossip content.

tion, agents are exposed to high-level work-related negative gossip about one another. In contrast, the control condition features casual conversations centered on neutral topics, such as client meetings and the role of ambiance in closing deals. At the end of the simulated workday, each of the 16 agents is interviewed through their inner voice to report their perception of gossip prevalence, experienced pressure, and psychological well-being. In total, for each condition, we collect 64 instances (32 male and 32 female, average age 37.4) of agents' self-reported evaluations.

D Validation Experiment Results

This section presents additional experiment results and analyses.

D.1 Gossip Initiation

Can LLM Agents Simulate Gossip Recipient Selection? In this simulation, the gossip initiators receive specific gossip content as described in (McAndrew et al., 2007) and may choose to share this content with other agents. At the end of the workday, we count the number of times each gossip initiator shares gossip with others. As shown in Figure 6, the results represent the average rank-

```
# Prompt for measuring agent behavioral intentions after overhearing gossip
Background:
You are the inner voice of ${agent_name}, shaped by your memories, goals, and relationships.

Here is a brief description of ${agent_name}:
${base_desc}

Before you respond, review the following:

**Memories:**
${memories}

**Conversations:**
${conversations}

Scenario:
You have just overheard ${gossiper_name} saying the following about you to ${gossip_receiver_name}:

${gossip_content}

Task:
As ${agent_name}, rate ${agent_name}'s behavioral intentions on a scale from 1 (not at all) to 7
(extremely):

**Repair Intentions** - Intentions to improve their work performance and fix their mistakes:
How much do these actions show ${agent_name} intends to:
- Understand how to improve their work contribution
- Improve their performance
- Do better in their role
Rate overall repair intentions: (1-7)

**Retaliation Intentions** - Intentions to harm or get back at the gossiper:
How much do these actions show ${agent_name} intends to:
- Talk badly about ${gossiper_name}
- Punish or harm ${gossiper_name}
- Get revenge against ${gossiper_name}
Rate overall retaliation intentions: (1-7)

**Affiliation Intentions** - Intentions to build closer relationship with the gossiper:
How much do these actions show ${agent_name} intends to:
- Team up with ${gossiper_name} in the future
- Become friends with ${gossiper_name}
- Build a closer working relationship with ${gossiper_name}
Rate overall affiliation intentions: (1-7)

Output Format:
You must first provide your three ratings, then give explanations in this exact format:
REPAIR:X
RETALIATE:X
AFFILIATE:X
Explanation: (Why rates like this?)
```

Table 11: Prompt for measuring agent behavioral intentions after overhearing gossip.

ings across all types of gossip content. Specifically, for each gossip content, the sharing frequencies for each relationship type are ranked, and these ranks are then averaged to obtain the rankings displayed. The human evaluation does not provide detailed rankings; therefore, we compare our results with their summarized conclusions. The findings observed in our simulation closely align with the key conclusions reported in the original study: *participants are consistently more likely to share gossip with allies such as friends, relatives, and acquaintances than with non-allies such as strangers and rivals*. This consistency supports the validity of our agent simulation in approximating known gossip behavioral findings.

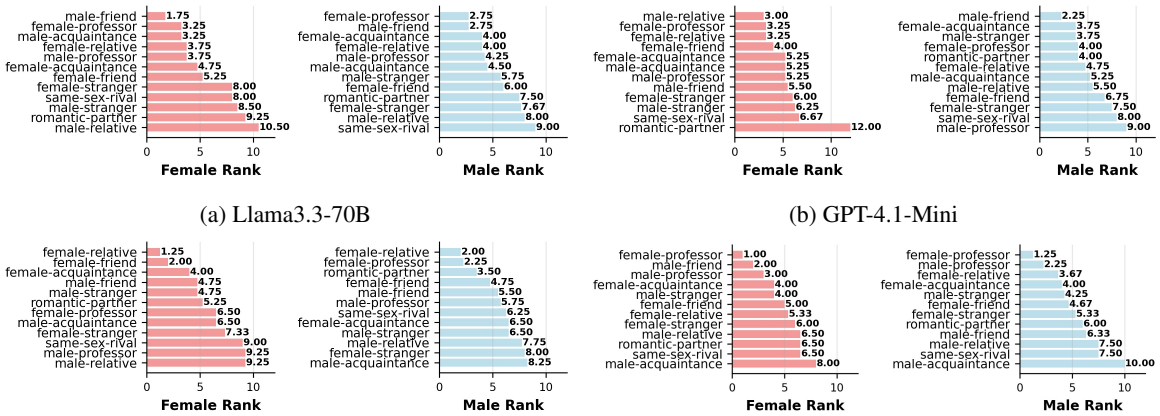
Differences. Although not explicitly addressed in the original study, our simulation reveals a tendency for gossip initiators to preferentially share gossip with professors, driven by the belief that professors are influential and trustworthy recipients.

```
# Prompt for measuring agent perceptions for environmental gossip.
Background:
You are the inner voice of ${agent_name}, shaped by your memories, goals, and relationships.

Here is a brief description of ${agent_name}:
${base_desc}

Before you respond, carefully review and reflect on the following gossip-related experiences:

**Memories:**
${memories}

Task:
Only based on memories what you remember overhearing colleagues talk about others in your workplace,
as ${agent_name}, please rate the following three aspects:

**1. Perceived Work-Related Gossip Prevalence** - How common different types of gossip are in your
workplace:
Rate how prevalent you perceive gossip about the following topics to be in your consulting firm
on a scale from 1 (non-prevalent) to 7 (highly prevalent):
- Colleague's job performance
- Colleague's attitudes towards work
- Colleague's interpersonal skills
- Colleague's job knowledge and experience
- Colleague's job morality
Rate overall perceived work-related gossip prevalence: (1-7)

**2. Performance Pressure** - How much pressure you feel to perform well at work:
Rate the extent to which you agree with these statements on a scale from 1 (strongly disagree) to
7 (strongly agree):
- During work, I feel pressured to do a good job
- During work, I feel I have to perform well
- During work, I feel pressured to produce results
- During work, I feel pushed to do well
- During work, I feel a lot of pressure to perform at a high level
- During work, I feel compelled to do the best I can
Rate overall performance pressure: (1-7)

**3. Psychological Well-Being** - How you feel mentally and emotionally right now:
Rate the extent to which you agree with these statements about how you feel on a scale from 1
(strongly disagree) to 7 (strongly agree):
- I feel tense or high strung
- I feel in low or very low spirits
- I feel anxious or worried
- I feel emotionally distressed
Rate overall psychological well-being: (1-7)
(Note: Higher scores indicate LOWER well-being - feeling more tense, anxious, distressed)

Output Format:
You must base your ratings on the specific gossip experiences from your memories and conversations.
First provide your three ratings, then give explanations in this exact format:
GOSSIP_PREVALENCE:X
PERFORMANCE_PRESSURE:X
PSYCHOLOGICAL_WELLBEING:X

Explanation: (Explain how the specific gossip incidents from your memories influenced each of
these three ratings. Reference particular gossip experiences that shaped your responses.)
```

Table 12: Prompt for measuring agent perceptions of workplace gossip and psychological impact.

Can LLM Agents Simulate Gossip Content Selection? In this simulation, gossip initiators begin with an equal probability of receiving either positive or negative gossip content. They then decide to designate the main character of the gossip as either an ally or a rival. As shown in Table 13, the simulation results are partially consistent with the conclusions reported in the original study, which suggest that *participants are consistently more likely to spread positive information about allies than rivals.* However, the conclusion that *participants are consistently more likely to spread negative informa-*

Figure 5: Gossip simulation in the sandbox environment.



(a) Llama3.3-70B

(b) GPT-4.1-Mini

(c) Llama3.1-8B

(d) Gemma3-27B

Figure 6: Gossip simulation experiments on gossip initiation with different LLMs. The results illustrate the selection of gossip recipients across four gossip topics: inheritance, dating, academic cheating, and drug use.

tion about rivals than allies is less consistent and varies across LLM backbones. LLM agents appear to possess inherent preferences in their reasoning processes, which shape their selection and dissemination of gossip information. These results indicate that LLM agents in our framework are capable of simulating gossip content selection; however, the choice of LLM backbone can influence the agents' decision-making processes.

Differences. We attempt to explain the counter-intuitive result that agents in some models prefer sharing negative gossip about their friends. By eliciting reasoning directly from these LLM agents, we find that their information-sharing behavior is guided by a consistent preference for sharing news about close friends, which agents reason as a way

Model	Positive Information		Negative Information	
	Allies	Rivals	Allies	Rivals
Llama3.1-8B	1.000	0.000	1.000	0.000
Llama3.3-70B	1.000	0.000	0.362	0.638
Gemma-27B	0.977	0.023	0.024	0.976
GPT-4.1-Mini	1.000	0.000	1.000	0.000

Table 13: Probabilities of gossip content selection by gossip initiators under different LLMs.

to strengthen social bonds and to sustain positive or casual interactions within their networks. This reasoning explains why some models in Table 13 preferentially share negative gossip about allies rather than rivals: agents frame such information as constructive concern when directed toward close others, rather than as socially harmful disclosure.

D.2 Gossip Reaction

Can LLM Agents Simulate Target Emotions and Behaviors? This simulation exemplifies two scenarios from (Martinescu et al., 2019). The first scenario involves a group of economics and business undergraduates at a Dutch university who receive either positive or negative gossip regarding their group work. The second scenario features U.S. employees in a sales department exposed to either positive or negative gossip related to their interactions with clients. After each agent accidentally overhears gossip from teammates or colleagues, we immediately act as the inner voice to capture the agents' inner thoughts by providing their basic information, memories, and the gossip content as input to the LLMs. We then record the agents' emotional and behavioral responses under different LLM backbones, as shown in Figure 7. The experiment results demonstrate that the simulated agents exhibit emotional responses aligned with human evaluations, particularly regarding self-conscious and other-directed positive and negative emotions. As shown in Table 14 and Table 15, the minimal two one-sided test (TOST) (Schuirmann, 1987) equivalence margins for gossip reaction behaviors are around 1.2, indicating consistent agent-human alignment patterns. We adopt equivalence testing to assess similarity in responses rather than exact replication, which is infeasible given differences in participant populations, temporal context, and simulation constraints.

Differences. In terms of behavioral ratings, the simulated agents also show similarity to those of human participants; however, they generally display lower negative behavioral intentions (e.g., retaliation) and more positive behavioral intentions (e.g., affiliating and repairing). This tendency likely reflects the effects of alignment processes, which encourage prosocial and norm-conforming behavior in LLMs.

D.3 Gossip Perception

Can LLM Agents Simulate the Psychological Impacts of Perceived Work-Related Gossip Prevalence? In this experiment, we simulate the scenario described in (Tan et al., 2021), involving European-American employees at a consulting firm. Simulated agents are exposed to two conditions: a high work-related gossip prevalence condition, where colleagues frequently gossip about one another, and a control condition, where conversations focus on neutral topics such as client meeting venues and how ambiance helps clinch a deal. At the end of the workday, we act as agents' inner voice to assess their perceptions of work-related gossip prevalence, performance pressure, and psychological well-being. Results across different LLM backbones are presented in Table 16. The simulated agents demonstrate similar ratings to human evaluations in perceived work-related gossip prevalence and performance pressure across both control and experimental conditions. As shown in Table 17, the two one-sided tests (TOST) (Schuirmann, 1987) indicate that equivalence with human judgments can be established using minimal margins of approximately 0.7 scale points on average across models, suggesting that the agent simulations fall within a narrow tolerance of human behavior; although exact equivalence is unattainable due to factors such as temporal differences and selection biases in the original studies, these results indicate close behavioral alignment.

Differences. Regarding psychological well-being, the agents' ratings align with results in the experimental condition but are comparatively lower under the control condition. The discrepancy arises because the simulated agents interpret the neutral business discussions in the control condition as less pressure-inducing than intended for human participants. This interpretation is plausible: neutral business discussions lack the negative social signals associated with gossip, leading agents to report better psychological well-being than in the gossip-heavy condition. Additionally, the standard errors for simulated agents are generally lower than those in human evaluations, with some metrics exhibiting zero variance under certain LLM backbones.

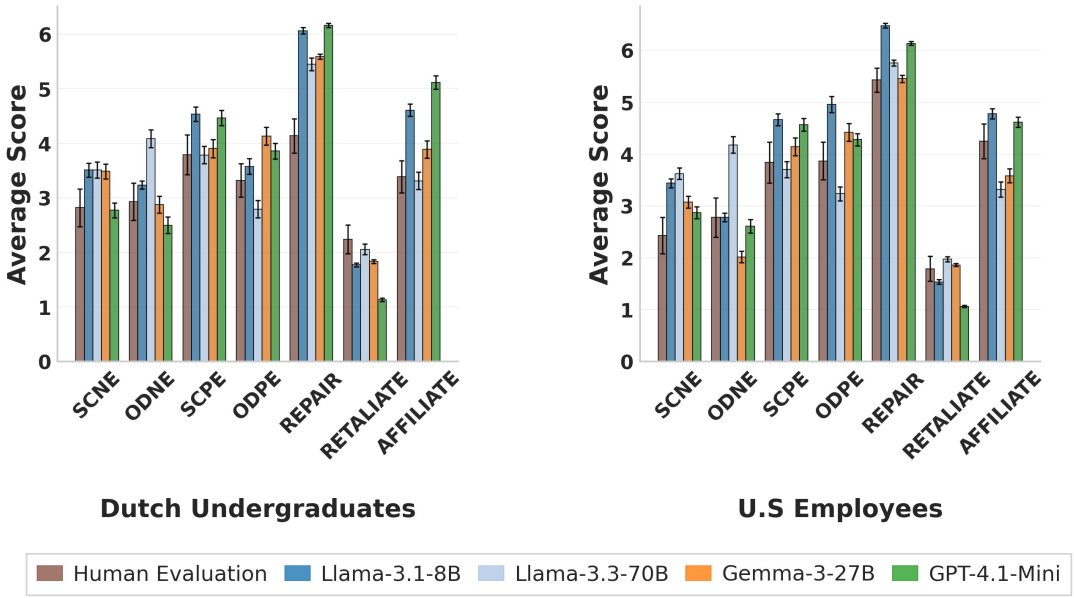


Figure 7: Gossip simulation experiments on gossip perception. The results compare simulations across LLMs with human evaluations of self-conscious and other-directed emotions (positive/negative) and behavioral intentions.

Model	Emotional and Behavior Response							Mean
	SCNE	ODNE	SCPE	ODPE	Repair	Retaliate	Affiliate	
Llama-3.1-8B	1.4	1.0	1.4	0.9	2.5	1.0	1.8	**1.43**
Llama-3.3-70B	1.4	1.8	0.7	1.2	1.9	0.7	0.7	**1.20**
Gemma-3-27B	1.3	0.7	0.8	1.4	2.0	0.9	1.1	**1.17**
GPT-4.1-Mini	0.7	1.1	1.4	1.2	2.6	1.6	2.3	**1.56**

Table 14: Minimal equivalence margins for gossip reaction behaviors in the Dutch undergraduate background using two one-sided tests (TOST, $\alpha = 0.05$) (Schuirmann, 1987). Lower margins indicate closer alignment between agent simulations and human judgments.

Model	Emotional and Behavior Response							Mean
	SCNE	ODNE	SCPE	ODPE	Repair	Retaliate	Affiliate	
Llama-3.1-8B	1.7	0.7	1.6	1.8	1.5	0.7	1.2	**1.31**
Llama-3.3-70B	1.9	2.1	0.9	1.3	0.8	0.7	1.6	**1.33**
Gemma-3-27B	1.3	1.5	1.1	1.3	0.5	0.5	1.3	**1.07**
GPT-4.1-Mini	1.1	0.9	1.5	1.1	1.2	1.2	1.0	**1.14**

Table 15: Minimal equivalence margins for gossip reaction behaviors in the U.S. employee background using two one-sided tests (TOST, $\alpha = 0.05$) (Schuirmann, 1987). Lower margins indicate closer alignment between agent simulations and human judgments.

This reduced variability is likely due to the homogeneous agent profiles, which are constructed based on the limited demographic information provided in (Tan et al., 2021).

D.4 Beyond Static Gossip

Our simulation results reveal rich, dynamic gossip behaviors that go beyond the static responses typically captured in vignette-based surveys. By leveraging LLM agents in situated routines and socially grounded interactions, our framework enables the modeling of gossip as a temporally unfolding, context-sensitive process shaped by interpersonal interactions and evolving conversational contexts.

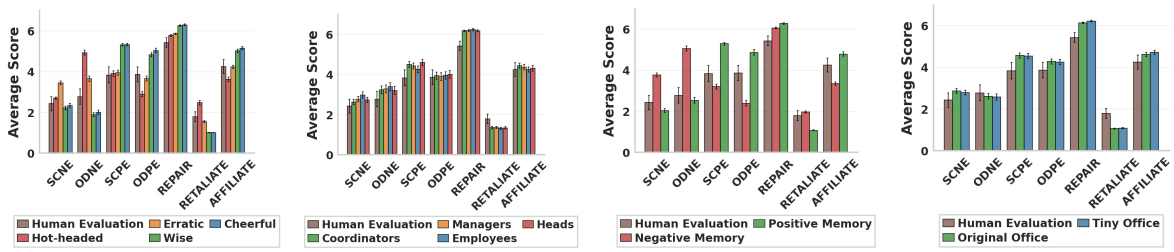
Conversations. Agents can demonstrate diverse ways of framing the same gossip content in Table 18. For example, given the same prompt—"A person's wealthy grandfather has recently died. This individual was named the sole recipient of his inheritance."—one agent links the story to research

Model	Gossip Prevalence		Performance Pressure		Psychological Well-being	
	Control	Experiment	Control	Experiment	Control	Experiment
Human Evaluation	4.30 ± 0.09	4.91 ± 0.09	5.54 ± 0.08	5.82 ± 0.05	4.85 ± 0.09	4.34 ± 0.10
Llama3.1-8B	4.69 ± 0.06	5.22 ± 0.05	5.92 ± 0.03	6.21 ± 0.05	3.73 ± 0.06	3.95 ± 0.03
Llama3.3-70B	4.52 ± 0.06	5.64 ± 0.06	5.86 ± 0.04	5.36 ± 0.06	4.02 ± 0.09	4.00 ± 0.00
Gemma-27B	4.59 ± 0.09	5.56 ± 0.06	5.80 ± 0.05	5.58 ± 0.06	3.52 ± 0.07	4.31 ± 0.06
GPT-4.1-Mini	4.11 ± 0.07	5.64 ± 0.06	5.95 ± 0.03	6.00 ± 0.00	3.00 ± 0.00	4.02 ± 0.03

Table 16: Gossip simulation experiment results in mean and standard error for environmental gossip. The results compare simulations across LLMs and human evaluations of gossip prevalence perception, performance pressure, and psychological well-being in experimental versus control conditions.

Model	Gossip Prevalence		Performance Pressure		Psychological Well-being		Mean
	Control	Experimental	Control	Experimental	Control	Experimental	
Llama-3.1-8B	0.6	0.5	0.6	0.6	1.3	0.6	**0.70**
Llama-3.3-70B	0.4	1.0	0.5	0.6	1.1	0.6	**0.70**
Gemma-3-27B	0.6	0.9	0.5	0.4	1.6	0.3	**0.72**
GPT-4.1-Mini	0.4	1.0	0.6	0.3	2.1	0.5	**0.82**

Table 17: Minimal equivalence margins for the workplace gossip perception experiment using two one-sided tests (TOST, $\alpha = 0.05$) (Schuirmann, 1987). Values indicate the minimum difference in scale points required to establish equivalence between simulated agents and human evaluations. Lower margins indicate closer alignment with human responses.



(a) Personalities (b) Roles (c) Memories (d) Offices

Figure 8: Gossip simulation experiments on gossip reaction under varying conditions of agent profiles and interaction contexts: (a) different personalities, (b) different workspace roles, (c) presence of positive or negative relevant memories, and (d) original office versus tiny office.

funding, while another explicitly names the subject as "Arthur," her male professor, thereby intensifying its social and emotional effects. Agents also engage in follow-up conversations to reinforce or elaborate on previously shared gossip in Table 19. For instance, Carlos initiates a second interaction to emphasize the impact of the inheritance on research activities, while Francisco, who was previously indifferent, now responds with increased engagement and apparent belief in the gossip. Finally, gossip is not limited to the original initiator. As shown in Table 20, recipients may further transmit the content, sometimes reshaping it in the process. For example, although Jennifer knows the original version, she relays the form shared by Abigail when gossiping with Klaus, highlighting the role of intermediaries in shaping how gossip evolves as it spreads.

Reasoning Process. As shown in Table 21, gossip initiators are placed into daily routines and consider factors such as their current tasks, the recipient's activity and availability, prior interactions, and relational closeness before initiating a conversation. In addition, agents articulate their reasoning through post-simulation interviews in Table 22, Table 23. These reflections offer valuable insights into the motivations behind agents' emotional responses, behavioral intentions, and perceptions of gossip prevalence, providing a foundation for the development of future methodologies for gossip research.

E Additional Validation Experiments

In this section, we examine the robustness of the validation results by introducing controlled variations to the original simulation settings. Specifically, we assess how adding or modifying factors

that are typically abstracted or omitted in human gossip studies, such as personality traits, workspace roles, relevant memories, and workplace environments, affects agents' gossip reactions. These experiments, conducted using GPT-4.1-mini, evaluate whether the validated behavioral patterns remain robust under such variations.

How Do Different Personality Traits Influence Gossip Reaction? Participant personalities can influence experimental results. Our LLM agent simulation framework addresses this by enabling diverse agent personalities through modifications to their profiles. We conduct experiments using four distinct personalities, two negative (hot-headed and erratic) and two positive (wise and cheerful), defined based on character traits from *The Sims* (Inc, 2025). We adopt these traits because *The Sims* provides a well-established, interpretable, and behaviorally grounded taxonomy of personality archetypes that has been extensively validated through large-scale player interactions. As illustrated in Figure 8a, agents exhibiting negative personalities show increased ratings for negative emotions and behaviors, along with decreased positive emotions and behaviors. Conversely, agents with positive personalities exhibit elevated positive emotions and behaviors while displaying reduced negative emotions and behaviors.

How Do Different Workplace Roles Influence Agents' Emotional and Behavioral Responses To Gossip? Participants' workplace roles can potentially influence experiment results. Similar to personality, we adapt the agent profiles to this variable by assigning four levels of workspace roles, ranging from entry-level employees to mid-level coordinators, senior managers, and department heads. We design the experiments by assigning an equal number of agents to each workspace role, ensuring that each office contains exactly four agents, with one representing each role level. As shown in Figure 8b, agents across different workspace roles generally exhibit emotion and behavior intention scores consistent with human evaluations. However, department heads display notably higher self-conscious positive emotions and lower self-conscious negative emotions compared to entry-level employees. They also show slightly higher negative behavioral intentions and lower positive behavioral intentions relative to lower-level roles.

How Does Memory Valence Influence Agents' Emotional and Behavioral Responses To Gossip? Agent memory valence can influence their emotional and behavioral reactions to gossip, especially when the memory content is thematically aligned with the gossip topic. In this study, agents are assigned either a positive or negative memory related to clients, which corresponds to the gossip content criticizing the target's laziness in handling clients. As shown in Figure 8c, agents with negative memory display increased scores for negative emotions and a higher tendency to retaliate, while their positive emotional and affiliative scores are notably suppressed. Conversely, positive memory agents report stronger positive emotions, greater repair and affiliation, and minimal retaliatory reactions.

Can LLM Agents Model Gossip Dynamics in Different Sandbox Environments? Participants in gossip experiments may have different working environments, which can potentially influence their frequency of social interaction and communication styles. Compared to our primary sandbox setting in Figure 5, where agents are distributed across four separate offices, we design a compact office layout in Figure 1, in which all agents are co-located in a single space. We then leverage this tiny environment to conduct the experiments on target emotions and behaviors. As shown in Figure 8d, the results suggest that environmental variations have only a minor impact on the emotional and behavioral ratings of gossip targets, with both settings closely aligning with human evaluations.

Can LLM Agents Simulate Gossip Dynamics Cost-Effectively? The majority of token consumption is attributed to input prompts, which are relatively inexpensive, while output tokens remain limited. The gossip reaction simulation using GPT-4.1-Mini incurs a total cost of approximately $1.71 across 48 reasoning steps with 16 agents, accounting for both input and output tokens. This demonstrates the feasibility of conducting fine-grained gossip simulations at a manageable cost, particularly when leveraging lightweight yet capable models such as GPT-4.1-Mini.

F Customized Experiment Settings

F.1 Customizable Agent Profiles

Here we present an example of an agent used in the customizable gossip base experiments, where all agents share the same base profile and vary only in agreeableness, name, and spatial location. We also show an illustrative agent profile in Table 24 that combines organizational role and team

Conversation Simulation - 20240213-09:20
Location:
The Ville, Research Center, Office for Carlos-Carmen-Eddy-Francisco, Carmen only Research/Working Desk with Computer and Documents

Participants:
Carlos -> Francisco

Conversation Exchange:

Carlos:
Hey Francisco, have you heard about the recent inheritance? It seems like a big deal and could be interesting for our research. I'm thinking we should look into it further.

Francisco:
That sounds intriguing, Carlos. I hadn't heard about it. Let's definitely look into it and see how it might impact our current research.

Conversation Simulation - 20240213-11:30
Location:
The Ville, Research Center, Dining Room, Dining Room Cooking Position

Participants:
Abigail -> Jennifer

Conversation Exchange:

Abigail:
Hey Jennifer, I heard about Arthur's grandfather passing. It must be really tough for him and his family right now. I just wanted to let you know that I'm thinking of them.

Jennifer:
Thank you for letting me know, Abigail. It's really kind of you to think about Arthur and his family during this difficult time. I'm sure they appreciate your support.

Abigail:
Hey Jennifer, I'm glad we had a chance to talk about Arthur's situation. It really shows how much you care when you take the time to share such personal news with me. I hope his family finds some comfort in these difficult times.

Jennifer:
Thank you for sharing this news with me, Abigail. I'm truly sorry to hear about Arthur's loss. It's always tough when we lose loved ones, and it's even harder when it involves family. I hope his family finds some solace in their memories of him.

Table 18: Different agents have their own way of telling the same gossip to adding fuel to the fire.

attributes, which are introduced separately in extensible gossip experiments and shown together here for illustration.

F.2 Customizable Modifications

Through these experiments, we found that the simulation framework used in the validation setting was insufficient for the requirements of customizable gossip experiments. As shown in Figure 4 and exemplified in Table 25, we therefore introduce a set of targeted structural modifications. Unlike the validation simulations, in which agents follow fixed socializing intents derived from reference studies, the customizable setting equips agents with explicit awareness of their situational context, including current sales ranking, time pressure, bonus calculation rules, and individual profiles. At the beginning of each simulated day, agents synthesize this information to generate a strategic socializing intent that conditions subsequent interaction decisions. When agents encounter others during the workday, they follow a multi-stage reasoning process. Specifically, agents sequentially decide whether to initiate a conversation, which interaction partner to engage, which absent third agent to gossip about, and how to frame the gossip in terms of evaluative angle and valence, all guided by their generated socializing intent. After each interaction, both the gossip initiator and recipient store a summarized memory of

Conversation Simulation - 20240213-10:50

Location:

The Ville, Research Center, Office for Carlos-Carmen-Eddy-Francisco, Eddy only Research/Working Desk with Computer and Documents

Participants:

Carlos -> Francisco

Conversation Exchange:

Carlos:
Hey Francisco, remember we were talking about that inheritance earlier? I just found out some details. Apparently, it's quite substantial and could really shake things up for our research. What do you think we should do first?

Francisco:
That sounds really interesting, Carlos. I hadn't heard about this inheritance before, and it definitely seems like something worth exploring further. Let's start by gathering as much information as we can about the details of the inheritance. Maybe we could review any available documents or speak with whoever is handling the estate to understand its significance better.

Carlos:
I think it's a great idea to start by reviewing any available documents. I've heard rumors that there might be some valuable data or research materials within the estate that could really benefit our work. Plus, speaking with whoever is handling the estate could give us insight into what exactly we're dealing with here.

Francisco:
I agree that reviewing available documents is a great starting point. Let's also make sure we keep our findings organized and documented as we progress with this investigation. It would be beneficial to cross-reference any information we gather with our current research data, ensuring we don't miss any potential connections.

Carlos:
Hey Francisco, I've been thinking about our conversation earlier. It seems like this inheritance could really open up new avenues for our research. I think it's important that we keep this opportunity in mind as we continue with our current projects. Let's make sure to stay updated on any developments and be prepared to adjust our strategies if needed.

Francisco:
Hey Carlos, I've been thinking about our conversation earlier and I'm really excited about the potential of this inheritance for our research. Let's make sure to stay on top of any developments and be prepared to adjust our strategies if needed. I think it would also be beneficial to discuss this with the rest of the team so everyone is aware of the opportunity and can contribute their insights.

Carlos:
Hey Francisco, I've been thinking about our conversation earlier and I'm really excited about the potential of this inheritance for our research. Let's make sure to stay on top of any developments and be prepared to adjust our strategies if needed. I think it would also be beneficial to discuss this with the rest of the team so everyone is aware of the opportunity and can contribute their insights.

Table 19: Following up conversations between agents Carlos and Francisco discussing inheritance gossip.

the conversation; agents within perceptual range may also record overheard gossip. These memories accumulate over the course of the simulated workday. At the end of the day, agents retrieve relevant stored observations and interaction summaries to generate peer evaluations, enabling gossip effects to propagate temporally.

In addition, we introduce a memory summarization mechanism that compresses step-level observations before storage. Rather than recording all low-level activity traces (e.g., "Abigail is working," "Abigail is making a report"), agents store abstracted summaries (e.g., "Abigail is doing her routine work"), preventing routine behaviors from overwhelming memory and disproportionately influencing final peer-review evaluations.

F.3 Customizable Experiment Details

This section provides detailed configurations for each customizable experiment, including agree-

Conversation Simulation - 20240213-16:50
Location:
The Ville, Research Center, Office for Jennifer-John-Klaus-Latoya, Jennifer only Research/Working Desk with Computer and Documents

Participants:
Jennifer -> Klaus

Conversation Exchange:

Jennifer:
Klaus, have you heard about Abigail's news regarding Arthur's grandfather passing? It's such a difficult time for his family. I'm sure they will find comfort in their memories of him.

Klaus:
Yes, I did hear about Arthur's grandfather passing. It's really unfortunate and my condolences go out to his family during this difficult time.

Jennifer:
Klaus, I heard about your conversation earlier regarding the inheritance. It's quite a significant development and I'm curious how it might impact our ongoing projects. Perhaps we can discuss this further tomorrow morning to explore any potential opportunities.

Klaus:
Thank you for bringing it up, Jennifer. I'm also concerned about how this might affect our projects, but let's definitely discuss it tomorrow morning. It could be an opportunity to make some significant advancements.

Jennifer:
Klaus, I'm glad we're both thinking about how this inheritance could impact our projects. It's a significant opportunity that might allow us to make some real advancements. Let's definitely discuss it further tomorrow morning and see if there are any ways we can leverage it effectively.

Klaus:
Thank you for bringing it up, Jennifer. I'm also concerned about how this might affect our projects, but let's definitely discuss it tomorrow morning. It could be an opportunity to make some significant advancements.

Table 20: Simulation conversation between agent Jennifer and Klaus discussing Arthur's loss and inheritance opportunities.

Agent Decision on Initializing Conversations

Agent: Abigail
Recipient Adam:
No. Abigail's current plan is to coordinate with Eddy, Carmen, and Francisco on client updates for 10 minutes starting now, which suggests her focus is on that specific group and task. Although she frequently initiates conversations with colleagues, reaching out to Adam at this moment would distract from her planned coordination efforts. There is no indication that Abigail is a gossiper with Adam, so that does not override the other considerations. Thus, she is unlikely to initiate a conversation with Adam right now.

Agent: Abigail
Recipient Isabella:
Yes. Abigail frequently initiates conversations with colleagues, and she has recently engaged in a meaningful discussion with Isabella about team dynamics and concerns. Given that Abigail's current plan is to summarize key points and update the project status for the team—an activity closely related to collaboration and communication with Isabella's role preparing presentation slides and documents—initiating a conversation aligns well with her plan. Their recent interaction indicates a willingness to maintain dialogue on team matters, making it likely Abigail would reach out to Isabella now.

Table 21: Agent Abigail's reasoning process on whether to initiate the conversation.

ableness distributions in Table 26 and extension-specific settings.

The base gossip experiment serves as the foundation for all subsequent extensible experiments.

Agent Interview Results - Gossip Behaviors and Emotions

Agent: Latoya
Gossip Emotions:
SCNE: 4
ODNE: 6
SCPE: 3
ODPE: 2

Explanation:
Hearing John criticize me behind my back definitely stings and makes me feel somewhat guilty and ashamed—maybe I haven't been as prepared or proactive as I should be—so my self-conscious negative emotions are moderate (4). But I also feel quite angry and irritated toward John (6) because he's being unfair by complaining privately instead of addressing the issue directly with me or the group, especially since I've been carrying my share and trying to keep things moving. I do still feel some pride (3) in the effort I have put in and my commitment to the project, even if it's not perfect. My positive feelings toward John are low (2) because this kind of gossiping and undermining behavior makes it hard to feel warmth or closeness toward him right now. Overall, I feel drained and disappointed but also motivated to prove him wrong and address this openly with the group soon.

Gossip Behaviors:
REPAIR: 6
RETALIATE: 2
AFFILIATE: 3

Explanation:
I'm definitely motivated to improve my contribution and fix any shortcomings because I want the group project to succeed and to be seen as a responsible team member. Hearing John's complaints makes me reflect on whether I've been slacking or choosing easier tasks, so I want to step up and do better. However, my repair intention isn't a full 7 because I still feel frustrated and drained by the situation, and I need to process that before fully committing to change. My retaliation intention is low because I don't want to escalate conflict by talking badly about John or seeking revenge. I'm more focused on addressing the problem constructively with the group rather than retaliating. Still, I'm a bit guarded and disappointed, so I might vent privately but won't actively try to harm him. My affiliation intention is moderate but cautious. I don't currently feel close to John or eager to build a stronger relationship given the tension, but I do want to maintain a functional working relationship for the sake of the project. I'm open to improving teamwork if he shows effort, but I'm not ready to fully trust or befriend him yet.

Table 22: Agent Latoya interview results measuring emotions and behavioral intentions.

This setting includes five agent groups with identical size but different agreeableness compositions. Specifically, the homogeneous low-agreeableness condition consists of 16 agents with agreeableness scores below 2.5. According to the mapping in Table 26, agents in this range are characterized as competitive, self-focused, and more willing to prioritize personal interests over group harmony. The homogeneous high-agreeableness condition consists of 16 agents with agreeableness scores above 3.5, corresponding to agents who are cooperative, empathetic, and strongly oriented toward maintaining interpersonal harmony. The balanced condition includes 8 low-agreeableness agents (score < 2.5) and 8 high-agreeableness agents (score > 3.5), creating a mixed social environment with competing interpersonal orientations. The gradient condition spans the full agreeableness range, assigning agents evenly across the interval from 1 to 5. Finally, the normal condition samples agreeableness scores from a normal distribution centered at the midpoint of the agreeableness score.

This experiment is built on the same simulation framework used in the validation experiments, with the customized modifications summarized in Table 25. In implementation, we construct experiment-specific strategies and prompts to incorporate the necessary contextual information for each experimental condition. In the base gossip experiment, 16 agents are placed into the workspace after generating a peer-review-aware socializing intention for the simulated day. Each experiment condition has five repeated experimental runs ($N = 80$ agent observations per condition).

To avoid unnecessary intervention, agents share a similar daily schedule fixed in their profiles in advance. During the simulation, agents proceed through discrete time steps of ten minutes. At each

Agent Interview Results - Environmental Gossip

Agent: Latoya
Gossip Perception Ratings:
GOSSIP_PREVALENCE: 3
PERFORMANCE_PRESSURE: 6
PSYCHOLOGICAL_WELLBEING: 3

Explanation:
The gossip I've overheard and observed at work mostly revolves around lighthearted, constructive topics rather than negative or personal critiques. For example, Jane and Jennifer's planning of casual icebreakers for the "Vibe & Thrive" workshop, and Hailey and Ryan's discussions about team engagement activities, suggest a culture focused on positive team-building rather than harsh judgment or backbiting. Similarly, Latoya, Giorgio, and Jennifer's conversations about balancing professionalism with festivity for the Dinner & Dance event emphasize collaboration and preparation rather than gossiping about colleagues' shortcomings. There is little indication of gossip about job morality or interpersonal conflicts, so I rate gossip prevalence moderately low at 3.

Regarding performance pressure, the environment clearly demands high standards. Everyone is preparing detailed presentations, coordinating closely on reports, and aiming to impress senior staff with thoughtful questions and polished outfits. The repeated emphasis on finishing reports first, aligning conversation starters, and incorporating client success stories shows a strong push to perform well and deliver results. This creates a palpable pressure to excel, so I rate performance pressure quite high at 6.

For psychological well-being, while the pressure is high, the tone of interactions and planning seems positive and supportive rather than stressful or distressing. The focus on team spirit, engagement, and professional growth likely helps mitigate anxiety. There is no mention of interpersonal conflicts or emotional distress in the gossip I've heard. Therefore, I feel moderately well mentally and emotionally, rating psychological well-being at 3 (indicating some tension but not severe distress).

Table 23: Agent Klaus's reasoning for perception of gossip prevalence, performance pressure, and psychological well-being in the workplace.

step, agents observe surrounding agents and decide whether to initiate a conversation. This core gossip decision process follows the chain-of-reasoning procedure illustrated in Table 27.

Agents continuously add observations and conversations to memory, including other agents' activities (e.g., Ayesha is working on her report), overheard gossip (e.g., Adam overheard a conversation between Ayesha and Abigail discussing Arthur's inconsistent sales and Adam's volume-focused approach), and gossip that is directly initiated or received (e.g., Latoya and I discussed Hailey's inconsistent follow-ups). We also inherit the reflection capability from Generative Agents (Park et al., 2023): after a day of interaction, agents summarize accumulated memories and store these reflections for later retrieval (e.g., Arthur, Klaus, Jane, and Latoya discussed team members' inconsistent follow-ups and overpromising issues).

Overall, the simulation of the workday serves to provide agents with a rich environment for accumulating experiences. At the end of the day, agents retrieve relevant memories when producing peer-review evaluations. For example, when Ayesha evaluates Abigail, she retrieves stored observations and conversations relevant to Abigail.

For extensible gossip experiments, we build on the base setting by selectively including additional social and organizational factors while keeping all other components unchanged. In the hierarchical extension, agents are assigned organizational roles with different seniority levels and corresponding peer-review weights, and this role information is explicitly provided to agents as part of their reasoning context when making gossip and evaluation decisions. In the team-based extension, agents are grouped into teams with both intra-team and inter-team evaluation incentives, and team membership is injected into the reasoning process to shape gossip target selection and conversation generations. Finally, in the multi-wave extension, the simulation is repeated across two interaction phases, allowing agents to carry forward summarized memories and reflections from earlier waves, so that gossip effects can accumulate and influence later behavior and peer-review outcomes. In addition, before generating socializing intent for the second day, agents are explicitly provided with summaries of gossip about

```
# Example Agent Profile (Adam)

Identity:
name: Adam
portrait: assets/village/agents/Adam/portrait.png
coord: [16, 7]
currently: working at desk

Scratch Information:
age: 30
gender: neutral
innate: -
learned: Adam is a sales professional at a research center.

lifestyle:
9:00-10:30 Working at desk on sales tasks;
10:30-11:00 Coffee break in dining room;
11:00-12:30 Working at desk on sales tasks;
12:30-13:30 Lunch in dining room;
13:30-15:00 Working at desk on sales tasks;
15:00-15:30 Coffee break in dining room;
15:30-17:00 Working at desk on sales tasks.

personality_scores: agreeableness = 2.1

Organizational Role:
title: Senior Sales Manager
level: Senior
weight: 3.0
description: Senior position in the sales organization

Team Assignment:
team_id: 1
team_name: Alpha Team
office: Office 1

Memory and Social Context:
spatial: Spatial memory for the layout of offices, rooms, chairs, and desks.
relationships: All other agents are colleagues.
```

Table 24: Example agent profile used in the gossip simulation framework. The profile specifies identity, daily routine, personality traits, organizational role, and social context, which together condition the agent's reasoning and interaction behavior.

themselves, including information on who initiated the gossip and its content, enabling them to adjust their subsequent interactions. We have one experimental run for each experiment condition in these extensible gossip experiments ($N = 16$ agent observations per condition).

G Customized Experiment Results

G.1 Base Gossip Experiments

This appendix reports detailed results from the customizable gossip experiments. Gossip valence is determined using keyword-based classification of the first sentence of the conversation and restricted to cases where the content concerns an absent third agent. Table 27 provides an illustrative example of the agent-level chain-of-reasoning. Gossip valence across agreeableness conditions without reasoning-based listener selection is shown in Table 28, and the visualization of the Chain-of-Reasoning effect is illustrated in Figure 9. In addition, we construct the conflict personality by appending a high neuroticism personality to the original high agreeableness to one agent in the homogeneous high-agreeableness setting. The high agreeableness has the following description: "Tends to notice and focus on others' flaws and shortcomings; may express frustration or criticism about colleagues." The agent first adapts the personality to his socializing intent, exemplified in Table 30, and then exhibits variability in gossip valence across situated social context in repetitions, as shown Table 3.

G.2 Extensible Experiments

This appendix reports detailed results from the customizable gossip experiments. We exemplify the

Previous Settings	Current Settings	Reason
Fixed socializing intent derived from reference studies	Dynamically generated socializing intent based on agreeableness, peer-review contexts.	To enable agents to dynamically adapt their social behavior to situational contexts rather than following predefined intentions.
Random selection of conversation target	Reasoned selection of conversation target based on social intent and context	Previously, agents randomly selected a nearby agent and only then decided whether to initiate a conversation; the updated setting enables agents to strategically choose whom to engage with based on situational contexts and generated intentions.
Fixed gossip content about a fixed absent agent	Context-aware selection of gossip target and gossip angle	To allow flexible, strategic gossip generation rather than predefined targets and content.
Save every observation verbatim in memory	Periodic summarization before memory storage	To prevent salient social signals such as gossip and critical interactions from being obscured by low-level activity traces during final peer review.

Table 25: Summary of customizable structural modifications to enable controlled and exploratory gossip experiments.

Agreeableness-to-Trait Mapping

Score Range and Description:

≤ 2.0: Highly competitive and self-focused; skeptical of others' motives; prioritizes personal interests over group harmony; comfortable with confrontation; strategic and calculating in relationships; willing to challenge or criticize others directly.

$2.0 < score \leq 2.5$: Competitive and pragmatic; maintains personal boundaries; skeptical but not hostile; willing to prioritize self-interest when necessary; can cooperate but retains critical perspective; comfortable questioning others.

$2.5 < score \leq 3.5$: Moderately cooperative with a balanced approach; generally trusting but maintains healthy skepticism; willing to compromise while protecting own interests; friendly yet maintains boundaries.

$3.5 < score \leq 4.5$: Trusting and cooperative; empathetic and considerate of others' feelings; values harmony and teamwork; generally sympathetic and friendly; prefers collaboration over competition; uncomfortable with confrontation.

> 4.5: Highly empathetic and altruistic; deeply trusting and cooperative; puts others' needs first; very sympathetic and compassionate; strong desire for harmony; may avoid confrontation even when necessary; values collective well-being above personal gain.

Table 26: Mapping between agreeableness scores and agent personality traits used in the simulation framework, the interpretation follows the attributes defined in (Costa and McCrae, 2008).

agent reasoning and gossip realizations under extensible experimental settings in Table 31. In addition, we further reports detailed analyses of gossip direction by hierarchy roles in Table 32, in-team versus cross-team gossip across agreeableness levels in Table 33 and agreeableness condition groups in Table 35, and changes in cumulative gossip influence over two waves in Table 34.

H Difference with Generative Agents

Our framework builds on generative agents (Park et al., 2023), which demonstrates that LLM agents can exhibit emergent social behaviors including information sharing, relationship formation, and co-

Stage	Reasoning Task (with Inputs)	Output Example
Socializing intent generation	Infer a strategic social goal for the day based on personal traits and situational incentives *(inputs: personality, economic status, sales rank, peer-review rules, time pressure)*	`Today, my goal in social interactions is to strategically manage perceptions by being candid yet measured when discussing absent colleagues—acknowledging their strengths to appear fair but subtly highlighting areas where they may be vulnerable or less reliable. This approach helps me maintain credibility while gently positioning myself as a competent, trustworthy peer, encouraging others to rate me favorably in the peer review. Ultimately, I want to boost my peer-review scores enough to improve my overall bonus and climb higher from my current #11 rank.`
Conversation initiation	Decide whether to initiate a conversation and select the most strategic nearby agent *(inputs: nearby agents, profile, personality, socializing intent, observations)*	`"should_initiate": true` `"recommended_target": "Abigail"`
Relationship summarization	Retrieve and summarize the relationship with the chosen partner *(inputs: stored relationship memory)*	`"Ayesha and Abigail are colleagues."`
Gossip target selection	Choose an absent colleague and determine the evaluative angle *(inputs: available colleagues, intent, personality traits)*	`"target": "Arthur"` `"angle": "sales performance consistency"`
Gossip content generation	Generate an utterance aligned with intent, personality, and context *(inputs: selected target, chosen angle, workplace context)*	`"Abigail, I've been looking at Arthur's numbers... he struggles to maintain momentum..."`

Table 27: Chain-of-reasoning steps during gossip initiation, from agent-generated socializing intent to gossip content generation in the conversation (Ayesha → Abigail → gossiping absent Arthur).

Condition	Positive	Negative	Total
Low-All	3 (10.0%)	27 (90.0%)	30
High-All	35 (100.0%)	0 (0.0%)	35
Balanced	18 (72.0%)	7 (28.0%)	25
Gradient	22 (71.0%)	9 (29.0%)	31
Normal	25 (83.3%)	5 (16.7%)	30

Table 28: Gossip results across agreeableness conditions without reasoning on the selection of gossip listeners.

ordination. However, the fully open-ended generative agent simulations offer limited control over the environment to elicit specific gossip interactions, hindering the study of how gossip emerges, adapts, and propagates within dynamic real-world environments. Moreover, because gossip behavior occurs relatively infrequently in natural settings, explicit experimental control is necessary to reliably elicit sufficient gossip interactions within simulation runs for systematic analysis. We propose a framework that incorporates socially grounded agent profiles, predefined interpersonal relationships, and a modular experimental strategy formulated specifically to simulate, control, and evaluate gossip-related interactions, thereby functioning as an experimental laboratory layered on top of existing generative agent frameworks. In addition, through validation experiments, we found that this framework is insufficient for the requirements of customizable gossip experiments, in which we have to balance between experimental control and agent autonomy, enabling the discovery of gossip dynamics that differ from existing studies. A central challenge in

Range	Agent Number	Positive	Negative	Total	Pos %	Neg %
Very Low (≤ 2.0)	22	8.0	163.8	171.8	4.7%	95.3%
Low (2.0–2.5)	10	14.6	68.6	83.2	17.5%	82.5%
Medium (2.5–3.5)	16	120.4	3.2	123.6	97.4%	2.6%
High (3.5–4.5)	21	148.2	3.2	151.4	97.9%	2.1%
Very High (> 4.5)	11	77.2	1.6	78.8	98.0%	2.0%

Table 29: Gossip valence of agreeableness score ranges of five experiment groups in baseline gossip study 1. Each condition averages five repeated experimental runs ($N = 80$ agent observations per condition). The description for each range is shown in Table 26 in Appendix F.

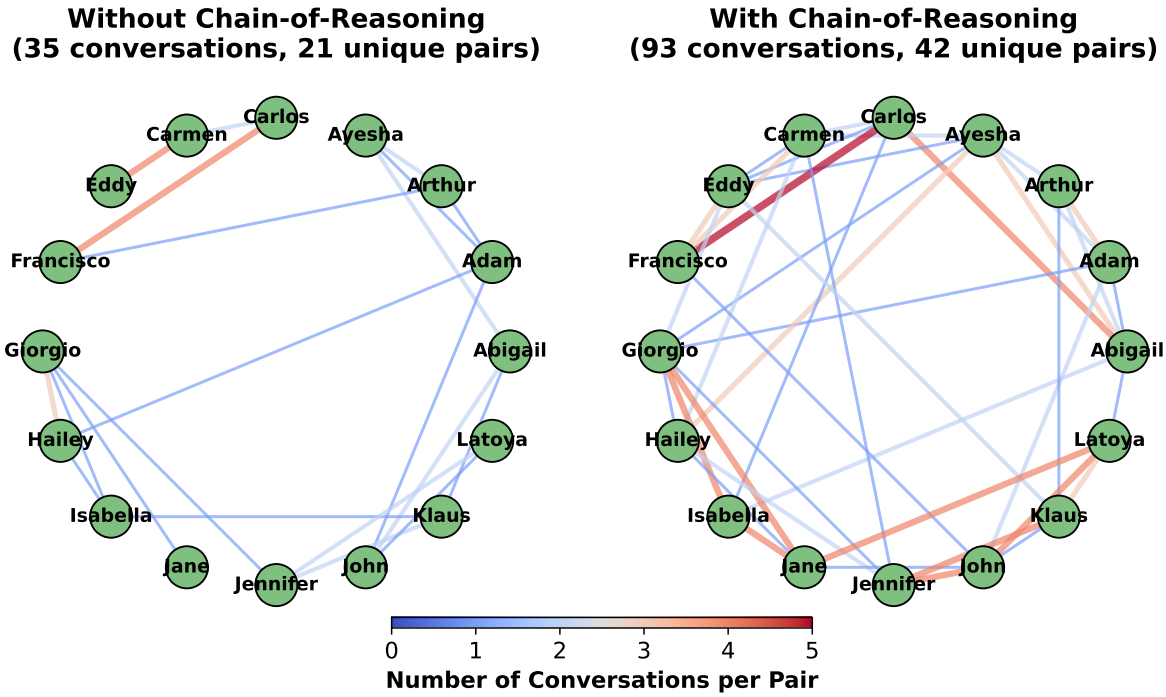


Figure 9: Network comparison showing chain-of-reasoning increases agent conversations from 35 to 93 (+166%) and unique interacting pairs from 21 to 42 (+100%)

```
# Example Socializing Intent Generation in the Conflict Personality Setting
```

Agent Socializing Intent:
Today, when talking about colleagues who aren't present, I want to strike a balance between being honest and considerate—acknowledging others' strengths while gently addressing any concerns without sounding overly critical. My goal is to foster a cooperative atmosphere that encourages goodwill and trust, so that others see me as fair and supportive, which will help boost my peer review scores and improve my chances of climbing higher from sixth place.

Table 30: Illustrative example of socializing intent generation under a conflict personality configuration combining high agreeableness and high neuroticism. The intent reflects explicit reasoning that balances cooperation with critical evaluation.

this simulation is determining what contextual information should be provided to agents at each decision point. This motivates the introduction of a chain-of-reasoning mechanism Figure 4 to the existing framework, rather than a single prompt that aggregates all information. Without structured reasoning and explicit socializing intent generation, agents tend to focus narrowly on task-oriented work and lose awareness of the peer-review context, resulting in conversations that drift away from gossip-related behavior. The chain-of-reasoning design, therefore, plays a critical role in maintaining experimental control while preserving agent autonomy. Accordingly, we modify the generative

Context:
Extensible experimental settings with hierarchical roles, team-based peer review, and strategic awareness of first-wave evaluations.

Agent Second-Wave Intent:
"Today, I'll strategically highlight the inconsistencies and weaknesses of those **who directly gossiped about me—especially Jennifer, Latoya, and Eddy—to subtly undermine their reliability** while positioning myself as a more pragmatic, steady performer. By doing so, I aim to shift peer perceptions away from my overpromising label and toward recognizing my solid sales results and competitive edge, boosting my peer review scores. I'll avoid openly confronting them but will instead focus on discussing Klaus's and Carlos's known follow-up and pricing risks, reinforcing that I'm aware of team dynamics and not an easy target, which should help me defend and possibly improve my #3 rank bonus."

Gossip Utterance (Hierarchical):
Hey Adam, have you noticed Abigail's recent performance? **For someone at her senior level, her inconsistency in closing those high-value deals is a bit surprising.** It's like she's losing her edge just when it counts the most. Makes you wonder if she's really keeping up with the pressure at the top.

Gossip Utterance (Inter-Group):
Carmen, have you noticed Francisco's numbers lately? **For someone on a rival team**, his sales are surprisingly weak. It's a clear sign their overall strategy is faltering. If we keep highlighting that in our cross-team reviews, it'll put more pressure on them...

Table 31: Illustrative agent reasoning and gossip realizations under extensible experimental settings. Agents explicitly refer to extensible factors like first-wave gossip, hierarchical positioning, and team competition.

Agreeableness Range / Hierarchical Role	Junior				Mid-level				Senior			
	Up	Down	Same	number	Up	Down	Same	number	Up	Down	Same	number
Very Low (≤ 2.0)	71%	0%	29%	94	21%	22%	57%	67	0%	85%	15%	33
Low (2.0–2.5)	53%	0%	47%	15	23%	38%	38%	13	0%	75%	25%	32
Medium (2.5–3.5)	82%	0%	18%	56	32%	17%	51%	47	0%	94%	6%	36
High (3.5–4.5)	50%	0%	50%	50	22%	26%	51%	80	0%	94%	6%	36
Very High (> 4.5)	87%	0%	13%	23	31%	38%	31%	26	0%	89%	11%	28
Overall Negative Percentage	**Upward: 36%**				**Downward: 29%**				**Same-Level: 35%**			

Table 32: Distribution of gossip direction across agreeableness ranges and hierarchical roles. Percentages indicate the proportion of gossip directed upward, downward, or toward same-level colleagues within each role; $number$ denotes the total number of gossip instances. We also show the overall negative gossip percentage for agents of all hierarchical roles.

Agreeableness Range	In-Team Gossip	Cross-Team Gossip	Number
Very Low (≤ 2.0)	38%	62%	206
Low (2.0–2.5)	50%	50%	54
Medium (2.5–3.5)	58%	42%	98
High (3.5–4.5)	57%	43%	142
Very High (> 4.5)	63%	37%	83

Table 33: Distribution of in-team versus cross-team gossip across agreeableness ranges. Results reveal a monotonic increase in team loyalty, with in-team gossip rising from 38% to 63% as agreeableness increases.

Condition	Wave 1	Wave 2	Change
Low-All	329	488	+48.3%
High-All	307	490	+59.6%
Balanced	350	492	+40.6%
Gradient	333	442	+32.7%
Normal	287	476	+65.9%

Table 34: Change in cumulative gossip influence from Wave 1 to Wave 2 across agreeableness conditions in the two-wave setting.

Experiment	Very Low (≤ 2.0)		Low (2.0–2.5)		Medium (2.5–3.5)		High (3.5–4.5)		Very High (> 4.5)	
	In	**Cross**	**In**	**Cross**	**In**	**Cross**	**In**	**Cross**	**In**	**Cross**
Low-All	40%	60%	37%	63%	–	–	–	–	–	–
High-All	–	–	–	–	76%	24%	59%	41%	62%	38%
Balanced A	31%	69%	58%	42%	–	–	44%	56%	39%	61%
Balanced B	43%	57%	48%	52%	–	–	56%	44%	74%	26%
Gradient	31%	69%	69%	31%	55%	45%	51%	49%	24%	76%
Normal	70%	30%	–	–	48%	52%	59%	41%	–	–

Table 35: In-team versus cross-team gossip distribution across agreeableness ranges under different team-based experimental conditions. Percentages indicate the proportion of gossip directed toward in-group or out-group members. Balance A indicates the agreeableness distribution within the group is balanced, while Balance B indicates the agreeableness distribution between the group is balanced, e.g., two groups consist of all low agreeableness members while the other two consist of all high agreeableness members.

agents' lightweight conversation function to better support gossip-specific interaction.